

Western University

Scholarship@Western

Digitized Theses

Digitized Special Collections

2009

GAUGING PUBLIC INTEREST FROM SERVER LOGS, SURVEYS AND INLINKS

Yijun Gao

Follow this and additional works at: <https://ir.lib.uwo.ca/digitizedtheses>

Recommended Citation

Gao, Yijun, "GAUGING PUBLIC INTEREST FROM SERVER LOGS, SURVEYS AND INLINKS" (2009). *Digitized Theses*. 4138.

<https://ir.lib.uwo.ca/digitizedtheses/4138>

This Thesis is brought to you for free and open access by the Digitized Special Collections at Scholarship@Western. It has been accepted for inclusion in Digitized Theses by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

**GAUGING PUBLIC INTEREST FROM SERVER LOGS,
SURVEYS AND INLINKS**

A Multi-Method Approach to Analyze News Websites

(Spine Title: Gauging Public Interest)

(Thesis Format: Monograph)

by

Yijun Gao

2

**Graduate Program
in Library and Information Science**

**A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy**

**The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada**

© Yijun Gao 2009

Abstract

As the World Wide Web (the Web) has turned into a full-fledged medium to disseminate news, it is very important for journalism and information science researchers to investigate how Web users access online news reports and how to interpret such usage patterns. This doctoral thesis collected and analyzed Web server log statistics, online surveys results, online reprints of the top 50 news reports, as well as external inlinks data of a leading comprehensive online newspaper (the *People's Daily Online*) in China, one of the biggest Web/information markets in today's world. The aim of the thesis was to explore various methods to gauge the public interest from a Webometrics perspective.

A total of 129 days of Web server log statistics, including the top 50 Chinese and English news stories with the highest daily pageview numbers, the comments attracted by these news items and the emailed frequencies of the same stories were collected from October 2007 to September 2008. These top 50 news items' positions on the Chinese and English homepages and the top 50 queries submitted to the website search engine of the *People's Daily Online* were also retrieved. Results of the two online surveys launched in March 2008 and March 2009 were collected after their respective closing dates. The external inlinks to the *People's Daily Online* were retrieved by Yahoo! (Chinese and English versions), and the online reprints were retrieved by Google.

Besides the general usage patterns identified from the top 50 news stories, this study, by conducting statistical tests on the data sets, also reveals the following findings. First, the editors' choices and the readers' favorites do not always match each other; thus content of news title is more important than its homepage position in attracting online visits. Second, the Chinese and English readers' interests in the same events are different. Third, the pageview numbers and comments posted to the news items reflect the unfavorable attitudes of the Chinese people toward the United States and Japan, which might offer us a method to investigate the public interest in some other issues or nations after necessary modifications.

More importantly, some publicly available data, such as the comments posted to the news stories and online survey results, further show that the pageview measure does reflect readers' interests/needs truthfully, as proved by the strong correlations between the top news reports and relevant top queries. The external inlinks to the news websites and the online reprints of the top news items help us examine readers' interests from other perspectives, as well as establish online profiles of the news websites. Such publicly accessible information could be an alternative data source for researchers to study readers' interests when the Web server log data are not available.

This doctoral thesis not only shows the usefulness of Web server log statistics, survey results, and other publicly accessible data in studying Web user's information needs, but also offers practical suggestions for online news sites to improve their contents and homepage designs. However, no single method can draw a complete picture of the online news readers' interests. The above mentioned research methodologies should be employed together, in order to make more comprehensive conclusions. Future research is especially needed to investigate the continuously rapid growth of the "Mobile News Readers," which poses both challenges and opportunities to the press industry in the 21st century.

Keywords: Webometrics, readers' interests, public opinion, Web server logs, surveys, inlinks, online newspaper, the *People's Daily Online*, China, news website, interlinking.

Acknowledgements

I am profoundly grateful to my chief supervisor Dr. Liwen Vaughan for her ceaseless support and guidance in completing this research. Thank you for pushing my thinking forward, for trusting my abilities and for your encouragement as I faced new challenges. It is tremendously fortunate for me to have worked closely with Dr. Vaughan, a true mentor offering me constructive advice from the very beginning of my doctoral studies.

Thank you as well to my supervisory committee members: Dr. Isola Ajiferuke and Professor Paul Benedetti for their insightful comments, assistance and close reading of my dissertation. I would also like to thank my former committee member Dr. Hong Cui for her contributions to the proposal of my thesis. All of these researchers formed an exceptional team that guided me through the thesis framing and writing process.

I am grateful to the faculty, staff, my colleagues and friends at the Faculty of Information and Media Studies at the University of Western Ontario, for their endless support as I completed my studies here. I would also like to express my gratitude to the staff members of the *People's Daily Online*, which provided me with the most valuable data. Without their cooperation, this project would have been impossible.

A special thank you to my parents (Mr. Shiqiang Gao and Mrs. Runhua Wu) and my sister (Miss Meijia Gao) who supported me to pursue my doctoral studies and inspired me to complete my research at Western. I also sincerely appreciate Mr. Jeffrey Malins at the University of Western Ontario for his important role in helping me edit my thesis.

Table of Contents

CERTIFICATE OF EXAMINATION.....	ii
Abstract.....	iii
Acknowledgements.....	v
Table of Contents.....	vi
List of Appendices.....	ix
List of Figures.....	x
List of Tables.....	xi
 Chapter 1 – Introduction and Background of the Study.....	 1
1.1 – Choosing the <i>People’s Daily Online</i> of China as a Research Subject.....	3
1.1.1 – China has the World’s Largest Online Population.....	3
1.1.2 – The <i>People’s Daily Online</i> ----China’s Major News Vendor.....	4
1.1.3 – Brief Introduction to the <i>People’s Daily Online</i>	5
1.1.4 – Other Considerations.....	7
1.2 – Significance of the Doctoral Thesis	7
1.3 – Structure of the Doctoral Thesis.....	10
 Chapter 2 – Problem Statements and Research Questions.....	 11
2.1 – How to Investigate Research Question One (RQ1).....	12
2.2 – Key Issues in Dealing with Research Question Two.....	15
2.2.1 – Hypothesis for RQ2.....	16
2.2.2 – Background Information on RQ2.....	16
2.2.3 – How to Investigate Research Question Two.....	18
2.2.4 – Significance of Answering Research Question Two.....	19
2.3 – How to Investigate Research Question Three.....	20
2.3.1 – Editors’ Choices of the Important News Reports.....	20
2.3.2 – Editors’ Use of Important Homepage Positions.....	21
2.3.3 – The Influences of Homepage Position on Pageview Numbers.....	22

2.4 – How to Investigate Research Question Four.....	23
2.4.1 – Co-existence of Chinese/English Stories on the Same Issues.....	23
2.4.2 – Chinese vs. English Readers.....	24
2.5 – How to Investigate Research Question Five.....	25
2.5.1 – Publicly Accessible Data of the News Websites.....	26
2.5.2 – Studying the Publicly Accessible Data.....	27
Chapter 3 – Literature Review.....	28
3.1 – General Webometrics Theory Research.....	28
3.2 – General Website (Non-Search Engine) Server Log Studies.....	31
3.3 – Search Engine Query Log Studies.....	36
3.4 – External Inlink Studies.....	41
3.5 – News Website Studies.....	44
Chapter 4 – Data Collection Methodology.....	50
4.1 – Collecting Web Server Log Data.....	50
4.1.1 – Server Log Statistics Generated by WebTrends.....	51
4.1.2 – Time Span of the Collected Data.....	54
4.1.3 – Classification of the Top 50 Chinese/English News Reports.....	57
4.1.4 – Classification of the Top 50 Chinese/English Queries.....	58
4.1.5 – Choosing and Classifying Reports on the United States and Japan.....	60
4.1.5.1 – Collecting U.S./Japan-Related Stories.....	62
4.1.5.2 – Collecting Comments Posted to U.S./Japan-Related Reports.....	63
4.1.5.3 – Collecting Russia-Related News Reports and Comments.....	64
4.1.6 – Collecting Homepage Position Data for the Top News Reports.....	65
4.1.7 – Pairing the Chinese and English Reports on the Same Events.....	67
4.2 – Collecting and Classifying Online Survey Results.....	68
4.3 – Collecting and Classifying External Inlinks.....	69
4.3.1 – Choosing Appropriate Search Engines to Retrieve Inlinks.....	70
4.3.2 – Using Proper Queries to Retrieve Inlinks.....	70
4.3.3 – Classifying the Retrieved Inlinks.....	73
4.3.4 – Collecting Interlinking Data among News Websites.....	75
4.4 – Collecting Online Reprint Data.....	76

Chapter 5 – Data Analysis and Results.....	78
5.1 – Analyzing Web Server Log Statistics.....	78
5.1.1 – Learning Readers’ General Interests from Server Logs (RQ1).....	78
5.1.1.1 – Distribution Pattern of Top 50 Reports and Queries.....	79
5.1.1.2 – Examining the Top 50 Chinese/English Reports.....	80
5.1.1.3 – Examining the Top 50 Chinese/English Queries.....	88
5.1.1.4 – Top News Reports’ Pageview Numbers vs. Comment Counts.....	90
5.1.2 – Chinese Readers’ Interests in U.S./Japan-Related Events (RQ2).....	91
5.1.3 – Top Reports’ Pageviews and Their Homepage Positions (RQ3).....	97
5.1.4 – Chinese and English Readers’ Interests (RQ4).....	100
5.1.5 – Exploring IP Address Distribution.....	102
5.1.6 – Traffic Graph to the <i>People’s Daily Online</i> by Alexa.....	105
5.2 – Online Survey Results Analysis.....	108
5.2.1 – The 2008 NPC Online Survey Results.....	108
5.2.2 – The 2009 NPC Online Survey Results.....	110
5.2.3 – Server Log Statistics vs. Survey Results.....	111
5.2.4 – Survey 2008 and Survey 2009.....	112
5.3 – Analyzing Inlinks to the <i>People’s Daily Online</i>	114
5.3.1 – Inlinks Profile of the <i>People’s Daily Online</i>	114
5.3.2 – Inlinks and Readers’ Interests.....	119
5.3.3 – Interlinking Structure among Leading News Websites.....	122
5.4 – Examining the Top 50 News Reports’ Online Reprints.....	126
 Chapter 6 – Conclusions and Future Studies.....	 130
6.1 – Findings of this Doctoral Thesis.....	131
6.1.1 – Chinese and English Readers had Different Interests.....	131
6.1.2 – Web Interaction Reflects Public Attitude on Specific Nations.....	134
6.1.3 – Editors’ Choices DO NOT Match Readers’ Needs.....	134
6.1.4 – Publicly Accessible Data of the News Websites.....	136
6.1.5 – The Inlinking Phenomena among News Websites.....	138

6.2 – Recommendations Made by the Thesis.....	139
6.3 – Limitations of the Study.....	141
6.4 – China’s Future Web and the Need for More Research.....	143
References.....	148
VITA.....	181

List of Appendices

Appendix 1: Classification Scheme for the Top 50 News Reports	159
Appendix 2: Classification Scheme for the Top 50 Queries.....	163
Appendix 3: 30 Negative News Reports on U.S.-Related Issues.....	168
Appendix 4: 30 Non-Negative News Reports on U.S.-related Issues.....	170
Appendix 5: 30 Negative News Reports on Japan-Related Issues.....	172
Appendix 6: 30 Non-Negative News Reports on Japan-Related Issues.....	174
Appendix 7: 30 Negative News Reports on Russia-Related Issues.....	176
Appendix 8: 30 Non-Negative News Reports on Russia-Related Issues.....	178
Appendix 9: Type of Website.....	180

List of Figures

Figure 4-1: Homepage of the <i>People's Daily Online</i> (Chinese Edition).....	66
Figure 4-2: Homepage of the <i>People's Daily Online</i> (English Edition).....	67
Figure 5-1: Histogram for the Top 50 Chinese News Reports' Pageview Numbers	78
Figure 5-2: Histogram for the Top 50 English News Reports' Pageview Numbers.....	79
Figure 5-3: Histogram for the Top 50 Chinese Queries' Counts.....	88
Figure 5-4: Histogram for the Top 50 English Queries' Counts.....	89
Figure 5-5: Histogram for U.S.-Related Negative News Reports' Pageview Numbers...	92
Figure 5-6: Histogram for Japan-Related Negative News Reports' Pageview Numbers..	92
Figure 5-7: Histogram for Russia-Related Negative News Reports' Pageview Numbers.	92
Figure 5-8: Histogram for Pageview Numbers for Chinese Top Reports on Homepage Pic.....	97
Figure 5-9: Histogram for Pageview Numbers for English Top Reports on Homepage Pic.....	97
Figure 5-10: Readers' IP Address Distribution for the Chinese Edition.....	104
Figure 5-11: Readers' IP Address Distribution for the English Edition.....	104
Figure 5-12: Traffic Graph of the <i>People's Daily Online</i> Generated by Alexa.com.....	105

List of Tables

Table 4-1: Inlink Search Results of Yahoo.com and Google.com	70
Table 4-2: Inlink Search Results of Yahoo.com and Yahoo.cn.....	71
Table 5-1: Spearman's Rho Test Results.....	80
Table 5-2: Frequency Distribution of Top 50 Chinese Reports from 129 Days.....	82
Table 5-3: Frequency Distribution of Top 50 English News Stories from 129 Days.....	84
Table 5-4: Distribution of Top 50 Chinese Reports excluding August 2008.....	87
Table 5-5: Distribution of Top 50 English Reports excluding August 2008.....	88
Table 5-6: Distribution of Top 50 Chinese Queries from 129 Days.....	89
Table 5-7: Distribution of Top 50 English Queries from 129 Days.....	90
Table 5-8: Correlation between Pageviews and Comments/Emailed Frequencies.....	91
Table 5-9: Mann-Whitney Test Results for U.S.-Related News Reports.....	93
Table 5-10: Mann-Whitney Test Results for Japan-Related News Reports	93
Table 5-11: Mann-Whitney Test Results for Russia-related News Reports	93
Table 5-12: Comments Posted to U.S.-Related News Reports.....	94
Table 5-13: Comments Posted to Japan-Related News Reports.....	95
Table 5-14: Comments Posted to Russia-Related News Reports.....	95
Table 5-15: Kruskal-Wallis Test Results for Chinese Top Stories on Homepages.....	98
Table 5-16: Kruskal-Wallis Test Results for English Top Reports on Homepages.....	99
Table 5-17: Distribution of 2008 Survey Votes.....	108
Table 5-18: Merging the 2008 Survey Options belonging to the Same Category.....	109
Table 5-19: Distribution of Top 50 Chinese Reports in March 2008	109
Table 5-20: Distribution of 2009 Survey Votes.....	110

Table 5-21: Merging the 2009 Survey Options belonging to the Same Category.....	111
Table 5-22: Distribution of Top 50 Chinese Queries in March 2008	112
Table 5-23: 2008 and 2009 Survey Results.....	113
Table 5-24: Originating Countries of the Inlinks.....	115
Table 5-25: Contents of the Inlinked Pages.....	116
Table 5-26: Purposes of Creating Inlinks.....	117
Table 5-27: Site Type of Inlinking Pages.....	119
Table 5-28: Queries to Collect Inlinks to the Chinese/English News Channel	121
Table 5-29: Inlink Counts to the Chinese/English News Channels.....	121
Table 5-30: Relations between Inlink Counts and Channel Pageview Numbers	122
Table 5-31: Interlinks among leading U.S. online newspapers.....	122
Table 5-32: Interlinks between <i>People's Daily Online</i> and its Chinese Partners.....	122
Table 5-33: Interlink Counts between <i>People's Daily Online</i> and U.S. Newspapers....	124
Table 5-34: Interlinking between <i>People's Daily Online</i> and U.S. Newspapers (1).....	125
Table 5-35: Interlinking between <i>People's Daily Online</i> and U.S. Newspapers (2).....	126
Table 5-36: Ties between Pageview Numbers and Online Reprints.....	127
Table 5-37: Pageviews vs. Online Reprints at the Group Level.....	128

Chapter 1 – Introduction and Background of the Study

The World Wide Web (the Web), which possesses far more than one trillion unique Uniform Resource Locators (Google, 2008), has turned into a full-fledged medium to disseminate news. Since the 1990s, thousands of traditional news providers around the world, such as newspapers, magazines, news agencies, as well as radio and TV stations, have rushed to establish their Web presence to attract more users and to increase revenue. Among them, online newspapers were characterized as “media of the future” (Thiel, 1998), and the number of newspaper websites around the world has at least doubled since 1999 (Feuilherade, 2004). Compared with their traditional print counterparts, these booming newspaper websites not only inherited the former's professional standards, but also added many remarkable advantages, such as immediacy, interactivity, unlimited space for multi-media reporting, hyperlinking, easy sharing, and more importantly, cost effectiveness. Therefore, an increasing number of people have opted for the Web to receive news or other information (Williams & Nicholas, 1999).

More importantly, harsh problems facing the traditional newspaper industry in the 21st century, such as falling circulation, soaring cost, delivery delays and waste generated by the daily printing system will be reduced or eliminated by newspapers migrating to the Web. In 2009, some renowned English newspapers (e.g., *the Seattle Post-intelligencer* and *the Christian Science Monitor*) even stopped the publication of their daily print editions and shifted most of their business to the Web so as to survive the ongoing economic downturn (Ngowi, 2009). The Central Daily News,

one of the world's oldest Chinese newspapers and the official "mouth-piece" of the Kuomintang (the ruling party of China before 1949, which now governs Taiwan) since 1928, also ceased its newspaper publication on June 1, 2006 and made the decision to dedicate its resources fully to online publication, due to mounting debts in the amount of \$30 million from the paper-based version (China Post, 2006).

How Web users access online news reports and what we can learn from such usage patterns is of significance to both journalists and information scientists. One of the methods by which these questions are addressed is the analysis of Web server logs (Thelwall, Vaughan & Bjorneborn, 2005). The greatest advantage of Web server logs is that such data show what people have actually done, and not what they say they might have done or thought they did (Nicholas et al., 2004). With the help of Web server log analysis techniques, it is easier now for researchers to monitor the general usage and traffic patterns of any news website, and then gauge the online newsreaders' interests, which is much more objective and reliable than results or answers from traditional questionnaires and surveys. Thus, findings from this type of study are crucial for the online newspaper's future development, and highlight the significant transitions of traditional information institutions in the 21st century.

In addition to the data generated by servers of the website being investigated, which is almost impossible to be retrieved by outsiders, some publicly available data, such as external inlinks and responses to online surveys, may offer us some supplementary clues in discovering the public's online interest. Bjorneborn and Ingwersen (2001) noted that "external inlinks" are hyperlinks pointing to a webpage

from sites outside the one being studied. More importantly, hyperlinks, one of the “defining features” of the Web, have some significant “social, political and even economic power” (Vaughan, 2005, p.949), and thus merit additional study in order to understand the hidden patterns of the newspaper website business.

1.1 – Choosing the *People's Daily Online* of China as a Research Subject

This doctoral thesis analyzed Web server log statistics, online surveys results, top news items’ online reprint rates, and external inlink data of *the People's Daily Online* (<http://www.people.com.cn>), a leading comprehensive news website in China, with the aim of gauging the public interest from the Webometrics perspective.

1.1.1 – China has the World's Largest Online Population

The number of Web users in China is now greater than the entire population of the United States, after rising to 338 million by the end of June 2009, according to the *24th Statistical Report on the Internet Development in China*. This report was released by the China Internet Network Information Center (CNNIC), a non-profit organization operated by the Chinese Academy of Sciences (CNNIC, 2009b).

China's online population, the largest in the world since June 2008, continued to grow by 40 million in the first half of 2009, CNNIC (2009b) said, adding that more than 90% of the online population (320 million) in China surf the Web via broadband access. The Internet penetration rate in China, however, stood at 22.6% by December 31, 2008, which was significantly lower than the U.S. rate of 71% (China Daily, 2008). The financial size of China's online market still trails that of the United States, South Korea and other countries (CNNIC, 2009a). All these data

indicate that there is much more room for continuously rapid growth of the Web and related business in China.

In recent years, the Chinese government launched a campaign to promote the knowledge economy by investing heavily in the information industry, and the number of websites in China amounted to 2.88 million by the end of 2008, 91.4% higher than the same period in 2007. The number of “.cn” domain names was 13.57 million, putting it ahead of “.de” (the country-code Top Level Domain for Germany), as the world's biggest country code (CNNIC 2009a). With all of these remarkable developments, China is taking shape as an influential Web power, and should not be ignored by the information science and journalism community.

1.1.2 – The *People's Daily Online* ---- China's Major News Vendor

The Web, dubbed “the fourth medium” following newspapers, radios and TV stations, was further eroding traditional media's turf in China by the end of 2008. More than 234 million Chinese people rely on the Web as their main source of news, and these “netizens” are generally educated and young (CNNIC, 2009a).

At present, all of the major players of the Chinese media industry have launched their own websites to provide the latest reports, such as *the People's Daily*, *the Xinhua News Agency*, *the China Central Television* and *the China Daily*. Altogether, only about 190 websites run by traditional media (e.g., the *People's Daily Online*) are licensed by the Chinese government as “exclusive news vendors” (Min, 2008). All other websites in China, big or small, can only reprint news items released by these 190 or so websites.

According to the first comprehensive national Web use survey conducted by the Chinese Academy of Social Sciences (CASS) in 2002, online readers in China praised the *People's Daily Online* as the "most trustworthy news resource" (CASS, 2002). The same survey also found that the Web has become a fast-growing information resource for Chinese netizens, and news browsing is one of their most important online interactions. CNNIC (2009b) further noted that most Chinese people used the Web for news, entertainment and social networking. Therefore, the *People's Daily Online* is a major news vendor that has contributed to the Web's development in China, and still plays an important role currently.

Due to the commercial sensitivity of the data, few studies have analyzed Chinese news websites' general usage patterns and readers' interests using Web server logs plus external inlink data, survey results, and online reprint rates of top news reports, thus leaving some knowledge gaps to be filled.

1.1.3 – Brief Introduction to the *People's Daily Online*

The *People's Daily*, which started publication in June 1948, has been ranked as one of the top ten comprehensive newspapers around the world by UNESCO. At present, the *People's Daily* has an average circulation of three million copies every day (including the domestic and overseas editions), and is the most influential and authoritative newspaper in China (People's Daily Online, 2009a). Its Web portal, the *People's Daily Online*, was launched on January 1, 1997, marking the *People's Daily* as a pioneer of Chinese online newspapers. Following this, the *People's Daily Online* turned into a full-fledged electronic information provider.

With more than a decade's efforts, the *People's Daily Online* has made itself one of the most authoritative, comprehensive and influential websites and a disseminator of information claiming the largest daily number of news releases in China (People's Daily Online, 2009a). Now it possesses websites in Chinese, English, Tibetan, Korean, Mongolian, Japanese, French, Spanish, Russian and Arabic: altogether ten language versions aimed at disseminating news and other information throughout the world.

The *People's Daily Online* Chinese Edition releases news around the clock with a daily updating of more than 1000 news pieces in various channels, including Chinese Politics, World Politics, Opinion, Business, Sci-Tech-Education, Society, Environment, Military Affairs, Entertainment, Life, Culture and so on, showing a far greater daily releasing capacity than its parental paper, which currently has only 20 pages per weekday.

The *People's Daily Online* English Edition (<http://english.people.com.cn>) has been accessible to the public since January 14, 1998. It has been making every effort to become a first-rate distribution center for information on China. Staff writers of this website translate major news releases from the *People's Daily*, as well as the latest policies or statements issued by the Chinese government, into English. Web users can also find useful information in special columns of the *People's Daily Online* about the activities of Chinese leaders, basic facts on China, and places of interest in China.

Nowadays, all reports published by the *People's Daily Online* since January 1997 (in Chinese) and January 1998 (in English) are archived and available for Web users to search and access, free of charge. Besides browsing current or archived reports, Web surfers visiting the *People's Daily Online* can post their comments on any news item in the online bulletin board following that report, or send URLs of the articles published by this site to friends' email boxes. These interactive activities contain much valuable information, and are also recorded by Web server logs.

1.1.4 – Other Considerations

The author possesses some unique advantages when conducting this research on the *People's Daily Online*. As a former staff journalist and freelancer with the *People's Daily* for more than ten years (since 1996), the author was one of the three founders of the *People's Daily Online* English Edition and served as senior editor of the Chinese Edition. There is no difficulty for the author to interpret the results of this study from the perspective of information science, Chinese culture and society, as well as open or hidden editorial rules for the Chinese news media. In October 2007, the *People's Daily Online* began to provide the author statistical data from the Web server logs for its Chinese and English sites for the purpose of a comprehensive readers' interests study, which is part of this doctoral thesis.

1.2 – Significance of the Doctoral Thesis

Web users in China are generally young and educated, CNNIC (2009a) noted, adding that people younger than 35 are the majority of online news readers, accounting for about 70% of total online population in China. Those readers'

interests and online activities may reflect some developing trends in Chinese society, and might represent substantial long-term social and economic impacts to this nation's progress in the near future. For example, the Web offers the younger generation an unprecedented way to express their views, and thus should not be ignored, because Web surfers in China represent the vital asset of the society. The *People's Daily Online* is deemed to be a cost-effective channel for the Chinese government to understand the public opinion of domestic and international affairs, which is important for China to make and adopt better policies.

There are numerous events happening and being reported everyday around the world; however, we cannot and need not cover all of them even with the help of the unlimited space availability of online news media. Therefore, all news agencies must grasp the usage patterns of their websites' contents in order to establish efficient and scientific principles to select and release news reports online, not only in accordance with their own editorial rules, but also satisfying readers' general interests/needs.

More than 1,000 journalists are dispatched to make exclusive contributions for the *People's Daily Online*, which also exchanges information with 350-plus cooperative news media agents. Thus, investigating the public's online interest becomes a much desired job to better meet the demands of the vast number of Web users. Adopting various methods to evaluate online newsreaders' interests will help the webmasters make strategic plans to improve their sites, such as providing more news of interest or having better solutions to make the readers' information seeking-

process much easier and faster (e.g., redesigning the homepage in accordance with the readers' online browsing behaviors). In this way, a news website may attract more visits, and in turn, more advertising clients or sponsors, which are crucial factors for an online news agency's survival and future development.

With continuous financial support from the Chinese government, the *People's Daily* will successfully survive the bleak economic winter (MacLeod, 2009). However, the delivery delay of this newspaper is still unsolved. For example, the *People's Daily* Overseas Edition always reaches its subscribers several days late. Here in London, Ontario, the paper is printed in Toronto, the biggest Canadian city some 200 km away, on the same day of its publication in Beijing. Sometimes, five days' worth of newspapers arrive together. Some Chinese subscribers, who have already read the same content from the *People's Daily Online* via the Internet instantaneously and free of charge, jokingly referred to it as "People's Weekly". Such a phenomenon calls for thorough study of online readers' interests, which is significant for the *People's Daily* to expand its business and reach in the 21st century.

The author maintains close working relationship with the *People's Daily Online*, and still serves as one of the top columnists for this online newspaper (See <http://world.people.com.cn/GB/guojj/209/7704/index.html> for details). Upon providing the needed server log statistics, the *People's Daily Online* required the author to submit a report on findings from this thesis, which were partially translated into Chinese by the author and aggregated into an internal report of *Development Strategy for the People's Daily Online* to improve its performance in

the coming years. The above mentioned findings include those results of relations between the top Chinese/English news reports' homepage positions and pageview numbers, the online readers' views on the United States and Japan, as well as the difference between the Chinese and the English readers' interests (Gao, 2009). Thus, by leveraging this leading online newspaper as the platform to conduct this interdisciplinary research, this doctoral thesis has made both theoretical and practical contributions in the fields of information science, journalism and studies of social phenomena.

1.3 – Structure of the Doctoral Thesis

- Introduction and general background information of the study (has been presented in the previous sections);
- Problem statements and research questions;
- Comprehensive review of related and current research in general Webometrics theory, Web server log analysis, inlink and survey explorations, as well as news website studies;
- Data collection, including server log statistics, survey results, top news items' reprint rates and external inlinks;
- Data analysis and results;
- Conclusions and future research;
- References and appendices.

Chapter 2 – Problem Statements and Research Questions

Understanding the interests of online newsreaders is very important for any news vendor. Given the large size of the Web audience and its diverse background, it is a difficult task for us to learn the online readers' interests through traditional methods alone (such as surveys or questionnaires). This thesis tried to explore the advantages and disadvantages of various methods to study the interests of online newsreaders, which will help us draw a better picture of an ever-growing and influential virtual community. Besides the issue of readers' interests, the navigation behaviors of the *People's Daily Online's* readers also merit study. However, such work is beyond the scope of the current thesis, and might be investigated by the author later. In one sentence, "learning what the readers need through various channels", is the focus of this doctoral thesis.

There are five research questions (RQ) to be addressed in this thesis:

RQ 1: What are the general interests of the *People's Daily Online's* readers?

RQ 2: What is the Chinese public interest in U.S. /Japan-related issues?

RQ 3: Is there any relationship between the top news reports' homepage positions and their pageview numbers?

RQ 4: Do Chinese and English readers have the same interests?

RQ 5: What can we learn from the publicly accessible data of news websites?

2.1 – How to Investigate Research Question One (RQ1)

The first research question explores which category of news reports published by the *People's Daily Online* attracted more visits (pageviews; see p.48 for details) during a one year period (from October 2007 to September 2008). Such information will help us grasp readers' general interests while browsing the *People's Daily Online*. This analysis was initially carried out by classifying the Chinese and English reports/articles published by the *People's Daily Online* in accordance with the following scheme:

- Chinese Politics (Chinese government's domestic or international policies and activities; other countries' policies or activities regarding China, such as choosing the new Communist Party and central government leadership; Taiwan and Tibet issues; Chinese military affairs; etc.);
- World Politics (foreign governments' domestic or international policies and activities, such as U.S. anti-terrorism wars in Iraq or Afghanistan; Russia's invading of Georgia during the Beijing Olympic Games; all military weapon developments; the U.N. Security Council's resolutions; etc.);
- Chinese Business issues (business/economic policies and performances, such as state-owned enterprises or rural economic reforms; China's stock and housing market issues; etc.);
- Business issues of Other Countries (U.S. financial crisis; the rocket rising of oil prices; world market/business information; etc.);

- Chinese Sci-Tech and Educational issues (the government's policies and achievements in the field of science, technology and education, such as China's launching of the Moon Orbiter and the first spacewalk);
- Sci-Tech and Educational issues of Other Countries (other governments' policies and achievements in the field of science, technology and education, such as Japan and India's moon exploration activities);
- Chinese Entertainment issues (e.g., Chinese movie or pop stars' activities);
- Entertainment issues of Other Countries;
- Chinese Sports issues;
- Sports issues of Other Countries;
- Chinese Health issues (the Chinese government's latest policies on health care reform; health, medicine or beauty tips; etc.);
- Health issues of Other Countries;
- Chinese Culture, Life and Society issues;
- Culture, Life and Society issues of Other Countries;
- Chinese Accidents and Disasters;
- Accidents and Disasters of Other Countries.

This classification scheme is based on the current news reports/articles classification scheme adopted by the parental newspaper of the *People's Daily Online*. Some minor changes have been made to maintain the consistency of the names (see Appendix 1 and Appendix 2 for the much more detailed classification scheme with Chinese/English report and query examples).

During the data analysis process, the classification scheme will be revised in accordance with the actual distribution of the top Chinese/English news stories, which includes the merging of some categories and creation of new ones. For example, Tibet or Taiwan issues are currently grouped into the Chinese Politics category; however, if reports on the two issues show up frequently, we need to consider adding the "Taiwan" and "Tibet" categories to group all of the Taiwan and Tibet-related news reports respectively. The Olympics will be another category made independent from the "Sports" category. Except for Chinese Politics and World Politics, all other "paired" categories will be merged to remove "Chinese" and "Other Countries". Reasons for such merging will be discussed in the data collection section (Chapter 4). Descriptive statistical (frequency) tests will be employed to address the first research question.

The above-mentioned relatively simple classification scheme was established also in line with the channel names of the *People's Daily Online* English Edition, which are far fewer in number than their Chinese counterparts. There are two channels' names that are not treated as news category names: Opinion and Photo. Articles and photos from these two channels will be grouped into other news categories in accordance with their contents, since "opinion", "photo" and "news" are the three categories of items from the *People's Daily Online* grouped by "type" rather than by "content". This revised classification scheme is compatible with both of the Chinese and English reports and reduces unnecessary complications during data collection and analysis.

2.2 – Key Issues in Dealing with Research Question Two

— *What is the Chinese public's interest in U.S. / Japan-related issues?*

In addition to learning the readers' general interests while browsing the *People's Daily Online's* news reports (i.e., which categories of news reports are popular), we could also conduct some studies of social phenomena with the help of modern Webometrics techniques, and explore whether or not the Web surfers' online interests could reflect the public opinion off-line. In this study, the author chooses to investigate the Chinese Web users' interests regarding the United States and Japan, which are not only China's two biggest trade partners but also political rivals and economic competitors.

Since the late 1970s, there has been no official policy or government-guided movements to encourage the anti-American and the anti-Japanese sentiments among the media or average Chinese people. On the contrary, the Chinese government always calls on its citizens to look to the future and decides to maintain close economic ties with the United States and Japan (China Daily, 2009; Xinhua, 2008a). The editorial rules of the *People's Daily* also require balanced reporting of international affairs. Will the Chinese people's online interest in the two countries be swayed accordingly? The author noticed that during the past few years, in the top news reports lists generated by the *People's Daily Online* (Chinese Edition), negative news reports on the United States and Japan were always ranked higher than their non-negative counterparts. However, we need a proper statistical test to address this issue.

2.2.1 – Hypothesis for RQ2

The hypothesis for the second research question is that the Chinese readers of the *People's Daily Online* tend to read more “negative reports” on United States/Japan-related issues than “non-negative stories”; in other words, “reports on human factor-involved negative events of the United States and Japan” attract more pageviews.

2.2.2 – Background Information on RQ2

There are anti-American and anti-Japanese sentiments among the average Chinese people nowadays due to historical and current causes. China and Japan had bitter relations from the mid-1890s till the end of the World War II (WWII), when Japan invaded China three times and killed millions of Chinese civilians. Since the 2000s, many Chinese people became furious with Japan after the former Japanese leaders' frequent visits to the Yasukuni Shrine, which enshrines the names of more than 1000 convicted WWII war criminals. Chinese citizens were also upset by some top Japanese officials' repeated rejections of the Nanjing Massacre by the Japanese Army in 1937, when 300,000 Chinese civilians and surrendered soldiers were killed in one month (Kahn, 2005; Encyclopedia-Britannica, 2009). China and Japan also have territory disputes over the Diaoyu Islands near Taiwan, as well as the oil fields on the East China Sea. Nowadays, many young and well-educated Chinese people still believe Japan is an “aggressor nation” and “should never be trusted” (Zakaria, 2009).

The Korean War and Vietnam War seriously tarnished the U.S. image in the minds of older Chinese people, who always called the United States "the Paper Tiger". Since the Former U.S. President Richard Nixon's historical visit to China in 1972, the two nations quickly moved closer, and established diplomatic relationships. However, the United States launched economic sanctions and a sensitive technologies embargo against China following the failed pro-democracy movement on Tiananmen Square in 1989. Anti-American sentiments have re-emerged in Chinese society since then.

In the 1990s, the U.S. Congress' unfriendly bills on China's human rights records, U.S.-China economic ties and Tibetan issues, as well as the U.S. government's selling of advanced weapons to Taiwan, repeatedly hurt the feelings of the Chinese people, and increased their anti-American sentiments. There were two big events that further fueled the anti-American sentiments among the Chinese people around the year 2000: the first was the U.S.-led NATO's bombing of the Chinese Embassy in Belgrade, capital of former Yugoslavia during the Kosovo War (People's Daily Online, 1999), which killed three Chinese journalists; the second was the collision of a Chinese military aircraft with a U.S. EP-3 spy plane near Hainan Island over the South China Sea, which killed the Chinese pilot (People's Daily Online, 2001). The unpopular Iraqi War also stimulated anti-American sentiments among the Chinese people, who maintain a negative perception of the U.S. government as "arrogant and imperialistic".

In 2008, especially after the unrest in Tibet's regional capital Lhasa, and the marred Olympic Torch Relay in Western Europe and the United States, there was strong nationalism and anti-Americanism among young Chinese people, who even called for anti-West boycott movements (Jacobs & Wang, 2008). A report by the U.S.-China Economic and Security Review Commission (USCC, 2005) also noted that there was anti-American sentiment in China. This thesis investigated whether such a phenomenon was still common among Chinese Web users during the time of this study.

2.2.3 – How to Investigate Research Question Two

The United States and Japan-related news stories published by the *People's Daily Online* can be divided into two groups. The first one includes reports on human error involving negative events (such as the financial crisis, all kinds of accidents, political failures or scandals, crimes, the attacks against or loss of the U.S. soldiers in Iraq/Afghanistan, etc.), and the second group of news items focusing on all other "non-negative issues" (such as cultural events, economic development, U.S. presidential elections, Sci-Tech progress, etc.). The Web server log data and proper statistical tests can tell us if "reports on human-related negative events of the United States and Japan" attract more visits than reports on non-negative events.

Considering online readers may actually like to access "negative news reports" (stories of negative events), conclusions made from the U.S./Japanese-related stories might be still vulnerable. Therefore, we need a "contrast/control group" (i.e., negative/ non-negative news reports on another country) to verify our

findings from examining United States/Japan-related stories. Due to the international status of the United States and Japan, we must also find an influential player on the world stage.

Compared with the strong anti-American and anti-Japanese sentiments among the average Chinese people in the past two decades, there is no obvious anti-Russian sentiment in China, although Beijing and Moscow verbally attacked each other and narrowly avoided a full-scale military conflict during the climax of the Cold War. After the collapse of the former Soviet Union, the Sino-Russian political, economic and military cooperation developed smoothly, and many Chinese people even hoped that China and Russia could form some kind of joint counter-measurement against the United States in the 21st century. More importantly, Russia, unlike other big powers in the West, never criticized China on its human rights records, Tibet (Dalai Lama) or Taiwan issues, significantly reducing the possible anti-Russian sentiment that could have arisen from the rare disputes between the two countries. Therefore, Russia is an appropriate contrast nation to help us investigate Research Question Two.

2.2.4 – Significance of Answering Research Question Two

Due to certain strict censorship regulations issued by the Central Publicity Department of the ruling Communist Party of China, "sensitive" news reports on China's domestic affairs, such as those articles on workers' strikes, conflicts between police and local people, demonstrations, serious accidents or natural disasters, spread of dangerous infectious diseases (e.g., the H1N1 Flu), and

important political and economic events, cannot be reported freely. News media from the Chinese mainland must publish/broadcast news items regarding the above-mentioned issues released by the *Xinhua News Agency* or the *People's Daily*. However, there are far fewer restrictions on reporting international events.

Such relative press freedom offers a good opportunity for us to investigate readers' interests in the United States and Japan, which may reflect the public opinion of the two nations, and make some contributions to the studies of social phenomena with the help of Webometrics techniques. Therefore, answering Research Question Two will pave the way for future projects, which may try to examine the Chinese people's impressions of any other country around the world through analysis of the Web server logs of the leading news websites in China.

2.3 – How to Investigate Research Question Three

-- Is there any relationship between the top news reports' homepage positions and their pageview numbers?

The hypothesis for RQ3 implies that the top news items' various positions on the homepage will yield significantly different pageview numbers; i.e., the editors' choices of important news reports will match the readers' favorites.

2.3.1 – Editors' Choices of Important News Reports

At the *People's Daily Online*, there is one well-known internal editorial rule: "put the most important reports in the most prominent places of the homepage to increase their visibility, and then to attract more readers", which is widely accepted by editors of this website. Thus, staff members with the *People's Daily Online's*

Chinese news channels not only update contents around the clock, but also try their best to compete for the so-called better homepage positions for the news reports from their own channels, since these employees are evaluated by the total number of visits to the news items/information from their channels. However, another editorial rule implies that the news titles of little or no political/economic significance cannot be put on the most obvious/visible sections of the homepage.

The "significant news reports" here refer to stories on the Chinese leaders' activities or speeches, policies of the Chinese government, and major international events. Sometimes, a few keynote editorials of the *People's Daily* or other opinion articles also enjoy such status and are published at the "best position" on the homepage of this online newspaper.

2.3.2 – Editors' Use of Important Homepage Positions

Opening the Chinese/English homepage of the *People's Daily Online*, all titles on the first screen focus on political/business issues of China and other nations, until the reader scrolls down the page to browse more titles. Titles here refer to the news headlines on the homepage without related pictures. The photo section of the homepage tells us another story: some much lighter topics, such as beauty pageants, new discoveries around the world, and culture/sports events will show up there instead of the tedious top leaders' "shaking hands ceremonies with foreign guests".

For the *People's Daily Online* Chinese Edition, two sections on the homepage are deemed to be the best: news title beneath the banner/channel name section (above all other titles) and homepage photos. Reports placed in the first position will

have their font size twice as large compared with any other titles listed on the homepage. Editors believe that the titles with larger font size and photos will help top news stories published there attract more visits than the plain text, regular-sized titles on all other sections of the homepage (see Figure 4-1 on p.63 for the layout of the Chinese homepage). Can the editors' hope become a reality?

For the *People's Daily Online* English Edition, the two sections for the most important news reports in the editors' minds are the Photo segment on the left and the large font sized titles segment on the right at the upper half of the homepage. Under these are titles considered to have "long-term interests to the readers", which are generally left there unchanged for more than three days, until something more important appears. News reports offered in this section include stories on the Communist Party of China's (CPC) top leaders' speeches and visits, or the government's policies. The daily updated news reports are listed at the lower half of the homepage in line with the time at which they were published (see Figure 4-2 on p.64 for the English homepage).

2.3.3 – The Influence of Homepage Position on Pageview Numbers

Since news titles with larger font size and the photo news items will never show up again in other sections of the homepage, we could try to figure out whether or not news titles placed in various sections of the homepage will generate significantly different volumes of online traffic as measured by pageview numbers. In other words, will the "larger font size title", "upper half of the homepage" or the "homepage photos" – rather than the contents of the news titles – make the top

stories attract more visits (pageviews)? Answering RQ3 will help us better understand the differences between "editors' choices" and "readers' needs" from the Chinese and English perspectives, which also reflect the online newsreaders' navigation behaviors, one important aspect of the information science studies.

2.4 – How to Investigate Research Question Four

--Do Chinese and English readers have the same interests?

In the past, Beijing relied on international broadcast, English newspapers, and journals to distribute China-related information, which had many limitations, such as delivery delays of printed material and the strict time schedule or special equipment requirements for the international broadcast services. Since the 1990s, the Information Office of the Chinese government has been asking its official English news websites to "make every effort" to become the "windows on China", and take the duties of releasing reports on the Chinese and world political/economic issues to English readers, so as to strengthen the "voice of China" (Wang, 2009; Xinhua, 2009).

2.4.1 – Co-existence of Chinese/English Stories on the Same Issues

The *People's Daily Online* offers English versions of its selected Chinese reports. Generally speaking, all the major Chinese/world political and economic events, latest Sci-tech progress and educational policies, entertainment snapshots, culture/life/society news stories, health care reports, sports news items, as well as keynote opinion articles are the focus of these English stories. These "paired" Chinese and English reports are of significance for editors from the two language

editions. The co-existence of Chinese and English news items on the same events raises an interesting question: will the top 50 English news reports get similar rankings compared with their Chinese counterparts? Such paired stories offer us another rare chance to study differences between readers' needs and editors' minds in the two languages.

According to the author's observations, in the top 50 news reports' lists, the rankings of the English news reports were always different from their Chinese counterparts on the same events. A statistical correlation test between the visit counts of the paired Chinese and English reports could uncover differences in online newsreaders' interests in the same events written in these two languages.

2.4.2 – Chinese vs. English Readers

Answering the fourth research question requires us to clarify the definition of the “Chinese” and “English” readers as well. According to the relative geographic distribution of the IP addresses generating visits to the *People's Daily Online* (Chinese Edition) during the data collection period, the majority of the IP addresses (approximately 85%) were limited to China (mainland). In the meantime, China (mainland) also hosted some 67% of the IP addresses that generated visits to the English Edition of the *People's Daily Online*. After this, the second-highest ranking country of the Chinese/English readers' IP address distribution was the United States.

The IP address distribution for the Chinese Edition was quite straightforward: most of these readers were Chinese speakers who live in China, as well as the rising number of Chinese speakers/researchers in the English world (i.e., the United

States). It is safe for us to assume that the vast majority of the English readers are foreigners living either in China or abroad. The assumptions regarding these two reader-populations are based on the following reasons:

- (1) There is an increasing number of foreign students/visitors/businessmen in China, who need information on China from the most authoritative channel;
- (2) The percentage (67%) of the IP addresses from China generating visits to the English edition is less than that visiting its Chinese counterpart (85%);
- (3) It is very unlikely that the average domestic Chinese person (other than researchers/English learners in China) would read news items from the English website run by a Chinese news agency, or the average English speaker (other than Chinese learners/researchers) would access news reports in Chinese.

2.5 – How to Investigate Research Question Five

-- What can we learn from the publicly accessible data of the news websites?

Few leading news websites would make their server log statistics, such as the pageview number of each report, accessible to the public due to their business sensitivities. Generally speaking, the server log statistical data from influential websites revealed to outside circles are the rankings of their top ten most viewed, most commented or most emailed news items, without indicating the exact pageview numbers or email frequencies associated with each story; an example is the “most popular lists” issued by the *New York Times*, the *Washington Post*, and *USA Today*.

2.5.1 – Publicly Accessible Data of the News Websites

Although server log data is not available to the public, there is much publicly accessible data we could retrieve from within and outside of these popular news sites. First, almost every news site nowadays encourages online visitors to interact with editors and other readers via the comment posting board thanks to the development of Web technologies. The number and contents of the comments are always available for other readers to browse, which may offer us some alternative methods to study readers' interests. This thesis will examine the relationships between the comment count and the pageview number of each of the top 50 news items, aiming to find whether or not such data will offer us valuable clues to uncover readers' interests. It is a good idea for us to thoroughly conduct a content analysis (qualitative study) on all comments on the top news stories. However, such interesting work is out of the reach of the current research project, and this doctoral thesis will only examine comments on the U.S./Japanese/Russian issues to determine public interest in them.

Second, there are two types of publicly available data sources for us to investigate the readers' general interests from outside of the online news media being studied: the external inlinks pointing to the news sites and the online reprint rates of their top news stories, thanks to the development of the commercial search engines Yahoo! and Google. Nowadays, the majority of websites in China, no matter their size, always re-publish "second-hand" news items within their own domains, which also attract lots of visits everyday and should not be ignored. It is illegal for

non-licensed websites to release news reports in China, so such websites depend on reprinting news items from licensed online media sources to feed their own news centers. These online reprints abide by copyright laws and use hyperlinks or plain text to indicate the sources of these reports.

2.5.2 – Studying the Publicly Accessible Data

The external inlinks, which could be treated as “citations” of re-published online news articles, as well as the online reprints themselves, may indirectly reveal readers’ interests not recorded by the server logs of the *People’s Daily Online*.

The author will try to investigate the relationships between the inlink counts/online reprint rates and the server log statistics. More interestingly, the domestic and foreign news agencies, especially major players, all create outgoing links to and receive inlinks from the *People’s Daily Online*. Such “interlinking phenomena” among news websites will also be explored in this thesis.

Answering the fifth research question will make some contributions to research methodology issues in addition to learning about readers’ interests by finding possible alternatives to using sensitive server log data.

Chapter 3 – Literature Review

3.1 – General Webometrics Theory Research

Tague-Sutcliffe (1992, p.1) noted that bibliometrics is “the study of the quantitative aspects of the production, dissemination and use of recorded information”; scientometrics is “the study of the quantitative aspects of science as a discipline or economic activity”; and informetrics is “the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists.” Since the mid-1990s, increasing efforts have been made to investigate the nature and properties of the Web as well as classic information retrieval systems by applying modern informetrics methodologies to their contents, space, structures and usage (Bjorneborn & Ingwersen, 2001; Wolfram, 2003). Such new quantitative studies on the construction and usage of the Web were named as “Webometrics” by Almind and Ingwersen (1997). As the Web becomes an ever more important medium for retrieving information, facilitating communication, and expanding business, more research is needed to map the nature of the user's interaction with the Web and to design better Web application systems, which provides us with fertile ground to explore new knowledge (Spink, 2002).

Web mining studies and Web technology analysis (i.e., search engine performance) are the two major areas of Webometrics research (Bjorneborn & Ingwersen, 2004), and the majority of such studies so far have focused on academic or scholarly Web use/communication (Thelwall, Vaughan & Bjorneborn,

2005). "Web mining" is a term first used by Etzioni (1996) to apply data mining techniques to uncover hidden features from Web documents/structures (links) and services, extract information from Web resources, and grasp general patterns in the Web or Web-related data (e.g., Web usage statistics from server logs). Web mining research can be further divided into three categories: Web content mining, Web structure mining, and Web usage mining (e.g., Kosala & Blockeel, 2000; Zhong, Liu & Yao, 2003). *Web content mining* aims to discover useful information from Web contents, such as text, image, audio, and video. It also includes resource discovery research on the Web, online document categorization and clustering (e.g., Kohonen et al., 2000), and information extraction from webpages (e.g., Hurst, 2001). *Web structure mining* focuses on building models underlying the hyperlink structures of the Web, which involves the analysis of inlink and outlink information of any website/page, and has been used for search engine result ranking as well as other Web applications (e.g., Brin & Page, 1998). *Web usage mining* employs data mining techniques to analyze server logs for the purpose of finding any interesting Web traffic patterns, which could help webmasters improve the performance of their sites (e.g., Lambert, 2008).

Most Webometrics research employs commercial search engines to retrieve quantitative data from the Web; thus the performance of popular commercial search engines is critical for the investigator to reach reliable conclusions, and as a result has attracted many scholars to investigate the issue. Thelwell (2008) employed 1587 single word queries to evaluate Google, Yahoo!, and Live Search.

This study recommended Google for hit count estimates and suggested Yahoo! for all other Webometrics purposes. It excluded Live Search for the data collection work of Webometrics studies, because this search engine returned significantly fewer results than the other two. This thesis uses Google to collect online reprints of the top 50 Chinese/English reports and uses Yahoo! to collect inlinks. Vaughan and Zhang (2007) examined four commercial search engines' (Google, Yahoo!, MSN, and Yahoo! China) coverage of 1,664 websites across the United States, China, Singapore, and Taiwan, within the commercial, educational, governmental, and organizational domains. Their study found that the U.S. domains and sites attracted more inlinks, which may give them a better chance of being indexed by the search engines. Vaughan and Zhang (2007) noted that Yahoo! China offered better coverage of sites from China or surrounding regions than the global Yahoo! in light of the remarkable development of the Web in China, which is an interesting issue related to the data collection section of this doctoral thesis (to be discussed in Chapter 4.3.2).

Based on the theoretical and empirical foundations established by previous Webometrics research, this thesis will employ Web intelligence mining technologies to explore the *People's Daily Online* of China. By analyzing this online newspaper's server log statistics and certain publicly accessible data, we can make some practical contributions to information science and media studies, since little research has been conducted on China's Web audience from the Webometrics perspective.

3.2 – General Website (Non-Search Engine) Server Log Studies

We are in an unprecedented position to examine the netizens' information needs and their use of online resources by analyzing the Web server logs that record online traffic data (Nicholas et al., 1999b). To investigate how online readers use one specific website, we can mine the Web server logs, which keep all requests sent to a server by users' browsers or Web crawlers/spiders/robots (Thelwall, Vaughan & Bjorneborn, 2005). The Web server logs contain valuable data that will lead to more informed decisions on the site's content and marketing strategies (Stone, 1999).

Cracking the Web server logs will help us genuinely meet the Web surfers' information needs, by unobtrusively tracking literally hundreds of millions of users around the world, far more than the number of traditional participants/subjects in the studies of information-seeking behaviors. However, these logs, which are breathtaking in their sheer volume and detail, and are of huge strategic and widespread economic worth, sometimes have been still described as "treasure troves of valuable data", because not all Web managers have the time and/or knowledge to interpret them (Nicholas et al., 1999a, p.263). Similar phenomena have also existed in China due to the business sensitivity of the server log data, as well as the above-mentioned factors; therefore, the *People's Daily Online* offered us a rare opportunity to have a glimpse at the landscape of the virtual community in China.

Numerous server log studies were done in past decades on how people interact with classic information retrieval systems; e.g., Thompson Dialog or library

Online Public Access Catalogs (OPACs), which were typically carried out in experimental settings to record activities of academic users, so as to improve user interfaces or training processes (Thelwall, Vaughan & Bjorneborn, 2005). However, the heterogeneous Web population has all kinds of interests/needs or reading habits while browsing the target site. For example, the online newspaper in the present study is updated much more frequently than any of the traditional information systems, which makes the lab-oriented research process non-applicable.

Fortunately, studies of the Web server logs from non-search engine websites (other than news websites) outside of the experimental settings have also been fruitful and uncovered much knowledge about usage of online electronic resources. These research projects reinforce the theoretical and practical frameworks for this doctoral thesis. For example, Wolfram and Xie (2002) investigated results of Web-based end-user survey and usage statistics of a digital library to find user characteristics, patterns of access and use, as well as user feedback. Cross-tabulations in this study used respondent demographics and revealed some key differences among various user groups: "older users valued the service more than younger users and engaged in different searching and viewing behaviors" (Wolfram & Xie, 2002, p.627). Collating data from their previous studies (e.g., Nicholas et al., 2000; Nicholas, Huntington & Williams, 2002) and triangulating with questionnaire data, Nicholas et al. (2004, p.24) summarized and characterized Web searching behaviour as "seldom penetrating a site to any depth, tending to visit a number of sites for any given information need, and seldom returning to sites they once

with online library collections. For example, Brown (2004) presented the methods and results of one year of tracking online federal document access through the University of Denver library OPAC. Xue (2004) evaluated a government publications library website by studying its usage statistics, examining access, searchability, and structure of electronic government information in the format of subject directories. Jana and Chatterjee (2004) assessed the accessibility and lineament of a website through statistical analysis of its site log files using hits, pageviews and user sessions as measurement, so as to identify some effective measures of usage and predict future usage using the linear trend line approach. This doctoral thesis also addresses the issues of readers' interests and online behavior characteristics by investigating the server log statistics collected for a one year period. The data collection software used by Xue (2004) and the pageview measure discussed by Jana and Chatterjee (2004) will be used in this doctoral thesis. Jacoby and Laskowski (2004) identified problems and issues in evaluation of electronic reserves and investigated usage measures, particularly Web server logs, for after-hours and off-campus access. They indicated that electronic reserves were heavily used during non-traditional service hours and from outside of the library's physical setting. Such an interesting finding promotes this thesis to examine whether the *People's Daily Online* receives more visits during the so-called "non-business hours". Cohen (2003) proposed a two-tiered model for analyzing website usage statistics for academic libraries: one tier indicating library usage patterns for institutional administrators; the other tier supporting server maintenance and site

design issues for website managers. Conclusions made by this doctoral thesis will also offer constructive advice to both the Editorial Board of the *People's Daily* and the administrator of the online newspaper.

Besides analyzing the server log statistics generated by the peripheral software, Nicholas et al. (2006b) applied "deep log analysis techniques" to study usage data on information-seeking behaviors of registered clients of Emerald Insight and Blackwell Synergy during the 12 months of 2002 and three months of 2003, respectively. By working with the raw server log data, these researchers revealed nearly three million users' information-seeking behaviors, such as "the extent to which they penetrate the site, the number of visits made, as well as the type of items and content they viewed" (Nicholas et al., 2006b, p.1345). Similar deep log analysis techniques were employed in the case study of Ohio LINK, a digital platform of some 6,000 full-text journals for more than a half million readers in Ohio, the United States, for the period of June 2004 to December 2004 (Huntington et al., 2006; Nicholas et al., 2006a). These two studies found the number and titles of the journals used, examined subject, date and method of access, and extracted sub-network and computer-label information from transactional server log files. More recently, Nicholas et al. (2008) analyzed the Web server (transaction) logs of some electronic journal libraries, and found a variety of patterns among scholars in their online full-text reading habits, such as the finding that many of the views were very brief in nature, since the articles might have been downloaded and read offline. The deep log analysis techniques

discussed here will not be used in this doctoral thesis; however, they shed some light on studying the *People's Daily Online* in the future.

The previously reviewed studies focus on Web server log analysis of various digital libraries, which are relatively static with fewer updating frequencies and visitors compared with news websites. This doctoral thesis, however, pays much attention to the wide-spread readers of one dynamic online newspaper, especially their interests in various news categories instead of the Web users' detailed browsing path within the domain of the *People's Daily Online*. The server log statistical data from this website is an extremely fertile ground for us to investigate the interests of the vast number of readers in real-world settings. Findings from this thesis are useful not only for designing better websites, but also for understanding certain social impacts of the Web, such as press freedom in virtual space or the public opinion on specific issues.

3.3 – Search Engine Query Log Studies

There are two types of search engines: general-purpose ones, such as Google and Yahoo!, and website search engines (such as the one employed by the *People's Daily Online*). The former sends Internet spiders/crawlers to automatically fetch webpages in order to create a vast index that can be searched by users (Chau & Chen, 2003), and does not restrict itself to particular domains or specialties. Although these general-purpose search engines have been trying to pool as many webpages as possible, there is no way for them to keep an up-to-date and complete search index, even as the number of indexed webpages by Google has exceeded one

trillion (Google, 2008). On the other hand, the website's own search engine allows users to seek any page within a specific domain. Such search engines can usually be found on almost every page of a website and only index pages from that particular domain. For example, visitors to the homepage of the *People's Daily Online* (<http://www.people.com.cn/>) will find its website search engine on the top of that page. Since the website search engine only needs to process a limited set of pages, its retrievable contents can be updated daily or even hourly, which is much more frequently than their general-purpose counterparts. Chau, Fang and Sheng (2005) pointed out that a website search engine always significantly out-performs general-purpose ones when retrieving information within the former's domain, because it can index all pages from the site, while no general-purpose search engine's crawler could cover every single page in a single site. Although website search engines are very useful for seeking information from one particular website, these search engines often have different user interfaces, interpret queries in different ways, support different types of advanced search functionalities, and even employ different search algorithms, which might complicate the end user's search experience (Chau, Fang & Sheng, 2005). More interestingly, Fagan (2002) analyzed an academic library website's search engine logs and found that patrons did not properly understand the functions of such a website search engine, and often tried to use it to find information other than the library's collections. Will a similar phenomenon occur for the website search engine of the *People's Daily Online*?

Despite the previously mentioned difficulties in server log analysis, much research has been done and important findings have emerged concerning the use of general-purpose search engines. Several commercial and research projects analyzed various search engines' query logs, and revealed valuable information on the online interests of users, which is also one of the major points to be discussed in this thesis. Studies on Excite's query logs were published widely (e.g., Ross & Wolfram, 2000; Spink et al., 2001; Wolfram, et al., 2001; Spink et al., 2002). AltaVista's query logs have also been analyzed by Silverstein et al. (1999). These studies pointed out that Web searching behaviors/interests by the public differ significantly from searching of traditional access-restricted/password-protected information retrieval systems (such as DIALOG and Lexis-Nexis) by their registered users. For example, most search engine users input short queries, seldom modify queries, rarely use advanced search features, and look at only the first 10-20 results listed by general search engines. Ross and Wolfram (2000) investigated the topics of queries submitted to the Excite search engine by classifying them according to content on the basis of term co-occurrence within unique queries, thereby providing a better understanding of the searched topics and their relationships. Spink et al. (2002) compared Excite server log data collected in September 1997, December 1999, and May 2001, and showed that search topics shifted but users' searching behaviors hardly changed.

In addition to the previous users' interests/behaviors oriented studies, Ajiferuke, Wolfram, and Famoye (2006) sampled queries submitted to the Excite search engine for informetrics characteristics, such as term frequencies, terms used

per query, pages viewed per query, and queries submitted per session, and generated various-sized data sets fitted to theoretical mathematical models in order to test the influence of sample size. These researchers then concluded that theoretical models do not satisfactorily fit data sets larger than 5000 observations. Wolfram (2008, p.1279) further analyzed transaction logs from four different Web-based information retrieval systems and noted that "users do engage in different search behaviors". Lewandowski (2008) examined the index freshness of the major Web search engines Google, Yahoo!, and MSN/Live.com through an examination of the updates of 40 daily-modified webpages and 30 irregularly-updated ones from a time span of six weeks in the years 2005, 2006 and 2007. He found that the best search engine in terms of "up-to-dateness" over the years was Google, followed by MSN and Yahoo!. More recently, Wolfram, Wang and Zhang (2009, p.896) applied cluster analysis to model session characteristics taken from large transaction logs of academic websites, public search engines and consumer health information portals, discovering "'hit and run' sessions on focused topics, relatively brief sessions on popular topics, and sustained sessions using obscure terms with greater query modification."

Considering the popular use of non-English Web search engines (i.e., Chinese ones) and their unique linguistic characteristics, Chau et al. (2007) analyzed three months' worth of queries from Timway (a search engine in Hong Kong) by examining the distribution patterns of search query terms. The study revealed that users submitted more diversified terms to Timway, and the usage of Chinese characters in the queries was significantly different from that of general online

Chinese contents (i.e., only a very small number of unique Chinese characters are used in search queries). Findings from the previously reviewed search engine studies all shed light on the current and follow-up research projects of this thesis.

More importantly, Bar-Ilan et al. (2009) designed an experiment to investigate user preferences for different rankings of search results from three major search engines (Google, Yahoo! and MSN Live), and noted that the users' choice of the "best" result from each of the different rankings indicated that "placement on the page (i.e., whether the result appears near the top) is the most important factor used in determining the quality of the result, not the actual content displayed in the top-10 snippets" (Bar-Ilan, et al., 2009, p.135). Results from this study are very relevant to one of the research questions in the current thesis regarding the homepage positions of top news reports and their pageview numbers.

Although the research mentioned above has greatly deepened our understanding on issues related to search engines and users' search behaviours, few studies have focused on the website search engines of online newspapers, especially those in Chinese (e.g., the one from the *People's Daily Online*), which may yield some promising results on readers' searching behaviours and interests, since the information needs of website search engines' users can be significantly different from those of the general-purpose commercial search engines such as Google (Wang, Berry & Yang, 2003). This doctoral thesis studied the query logs from the Chinese and English website search engines of the *People's Daily Online*, specifically to examine the differences in needs between two groups of readers.

3.4 – External InLink Studies

Examining the Web hyperlinks to the *People's Daily Online* is another focus of this doctoral thesis. Similar studies in the field of information science were initiated by Larson (1996), Rousseau (1997), as well as Almind and Ingwersen (1997), all using an analogy between citations and hyperlinks to explore new scholarly phenomena on the Web. Within the LIS community, parsing external inlinks pointing to webpages was widely accepted as reference and citation analysis of the scholarly communication on the Web: the traditional scientific entities (e.g., research articles and the citations) were replaced by webpages with incoming and outgoing links (e.g., Bjorneborn & Ingwersen, 2001; Wormell, 2001). Davenport and Cronin (2000, p.517) also proposed that Web links were “purveyors of trust in their targets,” which was proved by a series of significant correlations between Web link metrics and other measures in different contexts (e.g., Vaughan 2004b). The success of the link popularity algorithm pioneered by Google (Brin & Page, 1998) and similar link-based approaches adopted by other commercial search engines also made the conceptualization of Davenport and Cronin (2000) a promising one.

Cronin (2001) then stressed that hyperlinks should be treated as a new collective source of information on the Web, which were previously inaccessible or hard to quantify without the help of modern commercial search engines. Vaughan and Hysen (2002) found a significant correlation between the number of external inlinks and the Journal Impact Factor (JIF) for LIS journals: those with higher JIF

scores tend to attract more inlinks to their websites. The study also showed that the choice of search engine for data collection could affect the conclusion of a study; thus, multiple rounds of data collection is beneficial, especially when the result from a single round of data is borderline significant or inconclusive. In another research study on inlinks, Vaughan and Thelwall (2003) noted that both "longer site age" and "richer site content" are significant factors allowing the journals from library and information science and law to receive more inlinks. Meanwhile, one important measure of the visibility of a website is the number of inlinks that lead to the site, because "the more links to a site, the more chances the site will be visited" (Vaughan & Thelwall, 2003, p. 29). Payne and Thelwall (2007) examined the U.K., Australian and New Zealand academic websites from 2000 to 2005, and pointed out that the number of links and static pages in each of the three groups of academic websites seemed to have stabilized as far back as 2001. Such encouraging evidence shows that Webometrics research on academic spaces may have a longer-term validity than initially assumed.

Besides exploring the previously mentioned issues, studies of Web co-links were also popular in recent years. Vaughan (2006, p.1178) further examined Web co-links to Canadian university Web sites reflecting "the global view, the French Canada view, and the English Canada view", and applied Multidimensional scaling (MDS) to analyze and visualize co-linked data in a fashion similar to co-citation analysis. Mapping results of this study mirrored how Canadians viewed their universities and revealed linguistic and cultural differences in Canada. Vaughan,

Kipp and Gao (2007a, p.81) also conducted co-link studies on Canadian universities, and concluded Web co-links can be “a measure of the similarity or relatedness of sites being co-linked and that Web co-link analysis can thus be used to study relationships among linked Web sites”. Vaughan, Kipp and Gao (2007b) offered further evidence that co-linked telecommunications companies’ websites are really related through content analysis of 495 co-linking pages. These promising results show that co-linking is not a random phenomenon and that co-link data contain much information for Web data mining, especially in the area of investigating ties of different organizations or countries from the Web perspective (Vaughan, 2006).

In addition to the inlink analysis, many researchers also examine outlinks and interlinks among websites. Using data from top-level webpages across five high-level domains and from sample pages within individual websites, Ajiferuke and Wolfram (2004, p.43) studied the frequency distribution of outlinks in webpages, and paved the way for “simulation models of Web page structural content”. This study made it possible to estimate the number of outgoing links that might be found within webpages of a specific domain. Thelwall et al. (2008) assessed and graphed the Web connectivity data of European life science research groups, with the help of hyperlink-based techniques, such as commercial search engines and LexiURL, a special purpose link analysis tool. This study showed that Webometrics may offer us rich descriptive information on the international connectivity of research groups. However, little research has been conducted in the

area of online news sites' connectivity, which is an interesting issue to be addressed by this doctoral thesis.

Reid (2003) further noted that the interlinked nature of the Web will group together related sites and form implicit communities of one commercial site's internal and external stakeholders. Therefore, analyzing hyperlinks going to and coming from one site may not only help us better understand the Web, but may also help us track changes in society, technology, economy, and political environment (Fleisher & Bensoussan, 2003). A large number of studies (e.g., Thelwall, 2001; Vaughan, 2004a; Vaughan, 2004b; Vaughan & Hysen, 2003) have also suggested – by a series of significant correlations between Web link metrics and other measurement in different contexts – that Web links could be an indicator of the size or performance of an organization running a website in both the academic and commercial communities, thereby setting additional important theoretical and empirical frameworks for this doctoral thesis and future research.

3.5 – News Website Studies

Besides general Webometrics research, investigations of online news media have also blossomed, but few of these studies have a focus that overlaps significantly with this doctoral thesis. Williams and Nicholas (1999) surveyed U.S. and U.K. news websites and found that American newspapers were exploiting the advantages of Web information dissemination very well, such as hyperlinks, archived past reports, and reader interactivity. Massey and Levy (1999) analyzed the contents of 44 English-language online newspapers from 14 Asian countries, using a

five-dimensional conceptualization of interactivity. They noted that all of the online newspapers examined provided users with a relatively complex choice of news content, however, most did not rate highly on the remaining four interactivity dimensions. The *People's Daily Online* English Edition was not included in this research. Chyi and Lasorsa (1999) found national newspapers gained more ground online than local ones. Lin and Jeffres (2001) analyzed the contents of 422 online news media in 25 of the largest U.S. cities. They found that "each medium had a relatively distinctive content emphasis, while each attempts to utilize its website to maximize institutional goals" (Lin & Jeffres, 2001, p.555). Hope and Li (2004) surveyed hygiene and motivation factors affecting the quality of online newspapers. Hygiene factors here were essential requirements whose absence causes dissatisfaction (e.g., timeliness, content attractiveness, usefulness), while motivators were desirable elements that increase user satisfaction (e.g., writing style, layout, archives). Nicholas et al. (1999a, 1999b, 2000) conducted a series of research studies on the general reading interests of Web surfers browsing website of the *Times* and *Sunday Times*. AlShehri and Gunter (2002) employed an online survey to collect readership data and users' opinions of the electronic newspapers in the Arab world. Lewison (2002, 2003) even suggested mass media should be treated as an object of future informetrics research projects. This doctoral thesis chooses the *People's Daily Online* as a research subject, and also uses online survey to collect the needed data.

In addition to the previously reviewed news website studies, researchers also applied informetrics methods to analyze events (words) from online news media. Bar-Ilan (1997) focused on electronic news related to "*mad cow disease*", and revealed some similarities between the bibliometrics characteristics of online news items and those of print scientific works. Rousseau (2001) used search engines to track the occurrence of the words "*euro and euroland*" on the Web during 1999 when the euro was introduced as the common European currency. Chan et al. (2003) conducted a quantitative analysis of six newspapers' coverage of SARS, which examined the occurrence of the word SARS by collecting data from newspaper articles originating in Canada, mainland China, Hong Kong, and Western Europe.

The previously mentioned studies show the great potential informetrics tools have for analyzing the coverage issues of online news media, and more research in this direction has broadened the scope of Webometrics from covering mainly scholarly communication to general media (Chan et al., 2003). For example, Gao and Vaughan (2005) conducted an analysis of the nature of external inlinks (not the simple counts of numbers) to the *People's Daily Online*, *USA Today*, *The Globe and Mail*, and the *Singtao Daily* with the help of the search engine Yahoo!, and established Web hyperlink profiles of these newspaper sites. Part of the methodology from that study will be modified for this thesis. Massey (2004) shed some light on nonlinear storytelling on the Web editions of 38 U.S. daily newspapers and found that nonlinear storytelling, which employed unique features

of the Web (e.g., links to extra-content layers), was rare. Tremayne (2004) analyzed the use of hyperlinks in news stories on the Web. He examined about 1,500 online news reports over a five-year period, looking specifically at the hyperlinks embedded in these news items, and found that the use of links in Web news stories was increasing. But, this study did not examine where those links go.

In contrast to the previous two studies, which examined the quantity of links, this doctoral thesis examines both the quantitative and qualitative aspects of hyperlinks, especially in their important role as "online citations". Another difference is that the authors of these studies analyzed outlinks (links going out from the news websites to other sites) while the current study will investigate inlinks to the online newspaper. These collected hyperlinks' origination and destination places will also be examined. Tsui (2008) also studied the use of hyperlinks in news articles from online editions of four leading newspapers and five leading political blogs, which included the *New York Times*, the *Washington Post*, *USA Today*, and the *Los Angeles Times*, as well as the *Huffington Post*, *Michelle Malkin*, *Daily Kos*, *Crooks and Liars*, and *Think Progress*. The author noted that the U.S. newspapers were lacking in the use of outlinks pointing to resources outside of their domains. Similar interlinking phenomena are also investigated by this doctoral thesis: does the *People's Daily Online* use outgoing links to its Chinese and English counterparts or vice versa?

Compared with the previously mentioned online news media studies, there are not too many published reports relating to the newspaper website's log files and

their inlink analysis, even though the study of online surfing behavior has been very fertile (e.g., Tweddle et al., 1998), particularly in terms of analyzing the actions of users in a single site (Thelwall, 2001). Nicholas et al. (2000) focused on three months' worth of Web server logs of the *Times* and *Sunday Times* website, as well as a database of their subscribers, and successfully found the site readers' constitution, the time Web users spent online, and the pages people liked to read the most at the site. With that said, the last and the part deemed by Nicholas et al. (2000) to be the most interesting was rather simple, as the authors just indicated the news pages that were the most heavily requested, without mentioning what kind of specific subjects may have been popular. Hence, this doctoral thesis will try to draw a better picture of the public interest in China.

Wu and Bechtel (2002) also investigated the relationships between types of news events and daily traffic at the Web edition of *the New York Times*. Their results indicated that the level of disruptiveness and episodicity of the news stories were positively correlated with online traffic. Also, the number of reports on these topics of international politics, education, and science and technology, were positively correlated with online news usage statistics. However, the number of reports on domestic politics, weather, and accident and disaster news were negatively correlated with website visit counts at the time of the research. This doctoral thesis does not include similar research, and only focuses on which types of news reports attracted more visits, in order to gauge the online readers' general interests.

Development of the Web is changing the rules of news publishing and dissemination, which requires much radical re-thinking, especially for the newspaper industry. For example, He and Zhu (2002) examined the development of China's online newspapers (including the *People's Daily Online*) in respect to the social environment framework and noted several adverse factors, such as policies, regulations, economic structures, business conventions and telecommunication infrastructures. However, this online newspaper study did not mention usage patterns or readers' interests while visiting these news websites. Greer and Mensing (2004) conducted a content analysis of the same newspaper every winter from 1997 to 2003, and found trends in news presentation and content, multimedia use, interactivity and potential revenue resources, which will not be stressed in this doctoral thesis. D'haenens, Jankowski, and Heuvelman (2004) investigated how readers consumed and recalled news items presented in online and print versions of two newspapers in the Netherlands and found few differences between the online and print versions; in other words, the online readers' interests reflected those of the public. This doctoral thesis will not cover readers' usage of the print edition of the *People's Daily*, however; instead, it will investigate whether online Chinese readers' interests in United States/Japan/Russia-related news stories will mirror the public attitudes in the real world.

Chapter 4 – Data Collection Methodology

4.1 – Collecting Web Server Log Data

To study readers' interests while browsing *the People's Daily Online* or any other news website, researchers could sit behind real users or set up some video-cameras to capture how these subjects choose reports or conduct searches through its website search engine. However, such observations are not feasible for the large-scale evaluation of a website like the *People's Daily Online* and its search engine, because its millions of users are scattered across China and even around the world.

Although it is possible to capture users' actions (e.g., their clicking on a mouse button or scrolling a window on the screen) on their computers by using client-side monitoring techniques, it usually requires installing specific software or "plug-ins" on the users' computers. As these techniques require extra time and effort from users, and raise privacy and security concerns, most people do not want to install such unnecessary software and may change their natural reading habits under human or video surveillance (Chau, Fang & Sheng, 2005; Montgomery & Faloutsos, 2001).

Collecting server-side log data, however, can provide much information on netizens' information needs or interests, and is much more comprehensive and scalable than data retrieved client-side. This method requires no extra cooperation from Web browsers, and keeps their anonymity online. Such techniques can record every single visit to the target website or the website search engine being studied within any specific time slot (Zhong, Liu & Yao, 2003).

4.1.1 – Server Log Statistics Generated by WebTrends

There are many difficulties in analyzing Web server log files, such as identifying a user and determining a search session. In a traditional information retrieval system, users have to log in and out, so a user and a search session is explicitly identified. However, the “stateless” Web makes such identification extremely hard. The main measures that have been used to determine a unique Web user are cookies and Internet Protocol (IP) addresses. Silverstein et al. (1999) acknowledged that the cookie method is imperfect because different people can use the same browser – for example, a public computer station in a library – and users could disable the cookie functions of their browsers. Following one individual user’s search behavior over time within one site is much harder because this person can use different computers, such as one at home and one in the office. User identification by IP address is also imprecise because different users may share the same IP address in large ISPs and some organizations assign “floating IP address” to their computers. Besides, the server logs may not record the complete search process when caching, or proxy server techniques might be used by the visitors. It is also very difficult to determine users’ exact demographics, such as geographic locations. Although this kind of information can sometimes be obtained by an IP address reverse lookup software (e.g., the WebTrends Log Analyzer), such data could be misleading because many people may register their IP addresses in other countries for economic reasons (Thelwall, Vaughan & Bjorneborn, 2005). Hence, this thesis did not apply “visitors” or “sessions” as metrics, and the visits’

originating data obtained from IP addresses were treated as relative locations.

Both Tweddle et al. (1998) and Nicholas et al. (1999a) pointed out that proprietary Web server log analysis software may produce considerably different results due to the influence of “online noises” (e.g., Web crawlers' work; the graphics, banners, and menu frames on webpages; malicious hits by hackers or proxies), and suggested researchers use tailored spreadsheet packages to process the raw data. However, considering the sheer volume of the raw server logs, the author decided to use the statistical data generated by the WebTrends Log Analyzer installed on the server of the *People's Daily Online*, which efficiently helped to exclude visits generated by “online noises”, and has been employed by other studies (e.g., Xue, 2004). Using WebTrends Log Analyzer to generate the needed statistical data also significantly reduced the labor cost to the *People's Daily Online*, compared with asking its engineers to write and maintain their own program to process the raw logs.

For each selected day (from midnight to 23:59pm), the server log data pre-processed by customized WebTrends Log Analyzer was turned into human readable Web usage statistics, which include the following information saved in Excel files:

- The pageview number of each of the top 50 news reports published by the *People's Daily Online* (Chinese and English news items only);
- The number of comments posted to each of the top 50 news reports;
- How many times each of the top 50 news reports has been emailed to others;

- The top 50 queries submitted to the site search engine of the *People's Daily Online* (Chinese and English editions only);
- The relative geographic distribution of IP addresses generating visits to the *People's Daily Online* (Chinese and English news items only).

We employed the metric pageview (or page impression, as named by other researchers) for this study because it refers to a request to load a single page from a Web site. This concept is in contrast to the formerly popular "hit" metric, which refers to the request for a "file" (such as images, links, or frames on one single webpage) from a web server. Thus, there may be many hits per single pageview. For the Web, a page request would also result from a netizen's clicking on a link on another HTML page pointing to the page in question, which is the inlink issue studied in this thesis. For the current study, only an abnormally high volume of page requests (e.g., 1000) from a single IP address within a short time (e.g., 1 minute) were treated as "malicious" and removed from the statistics. Such IP addresses were also blocked by the *People's Daily Online*. While processing the raw server log data, WebTrend further filtered out internal traffic generated by the staff of the *People's Daily* and other "online noises", aiming to ensure that the data captured reflects the true usage patterns of the *People's Daily Online*.

During the initial stage of the data collection period (October 2007 and November 2007), the author collected the top 100 news stories from each selected date. By studying the pageview distribution pattern of the top 100 news items, it was observed that the majority of the total pageviews generated by the top 100 news

items (86.55% for the Chinese reports and 85.83% for the English ones) came from the top 50 news items. Thus, it was safe for us to only collect and analyze the top 50 news reports with the highest daily pageview numbers during the whole data collection period (from October 2007 to September 2008).

4.1.2 – Time Span of the Collected Data

For this doctoral thesis, server log statistical data for the *People's Daily Online* was provided by this news site from October 1, 2007 to September 30, 2008. One of the most important considerations for choosing such a time span for data collection is that many significant events were expected to happen even in the middle of 2007 when the author was revising the proposal:

- In October 2007, the 17th National Congress of the Communist Party of China's Central Committee chose the new leadership of the nation's ruling party;
- In March 2008, the National People's Congress named top leaders of the new Chinese government;
- In March 2008, Taiwan (a break-away province of China) elected its new leader;
- In August 2008, the 29th Olympic Games were held in Beijing.

Besides these “deemed to happen” events, many other breaking news stories emerged throughout the year of 2008:

- In January, unprecedented freezing rain/snow hit half of China's southern territory, causing havoc and severely disrupting the busy traffic before

China's most important family reunification: celebrating the Spring Festival;

- In March, riots broke out in Tibet's capital Lhasa and other remote cities;
- In April, the collision of two passenger trains took place in China, Hand-Foot-Mouth disease was spread in China's rural areas, and the Olympic Torch Relays were almost ruined in western Europe (e.g., France);
- In May, a massive earthquake struck southwest China's Sichuan Province;
- In June and July, some unrest happened across China (e.g., the protest in southwest China's Guizhou Province, and the killing of six policemen by one attacker in Shanghai);
- In September, the following events took place: the tainted baby formula scandal of China, the first Chinese spacewalk, and the on-going melting down of the U.S. financial market accompanying global recession fears.

These above-mentioned "hot events" not only stimulated more reports published by the *People's Daily Online*, but also attracted more visits to this website, which has claimed to have most comprehensive coverage of these breaking news stories, thanks to its almost unlimited Web space as well as its 24/7 on-duty professional editors and reporters. Therefore, traffic to this website reached a record high in September 2008.

More visits to the *People's Daily Online* during the data collection period may lead to more comprehensive conclusions on readers' interests. Although October 2007 and March 2008 are very good time periods to study the public's online

interest in China's domestic political/economic issues (e.g., foreign policy, finance policy, defense policy), they are still atypical months to draw conclusions on readers' interests in other issues, such as international events. Furthermore, what happened surrounding the Beijing Olympic Games of August 2008 also merits study; thus, server log statistics from August 2008 were also collected, along with the October 2007 and March 2008 data (the two months with many significant political events). In addition to that, the author further randomly sampled dates from the remaining nine months of the year, trying to make the collected data reflect the actual situation of visits to the *People's Daily Online*. In the meantime, monitoring the fluctuation of the traffic to the *People's Daily Online* between the "normal days" and the "days with breaking news" helped us to understand the changing of interests among online news readers within a 12-month-period full of all kinds of important events.

This thesis randomly chose one day from each week in the months other than October 2007, March 2008 and August 2008 from which to sample. In this manner, there was a total of 129 ($31 \times 3 + 36$) days' worth of Web server log statistics being retrieved and analyzed. To supplement these data, once breaking news happened (such as the Sichuan Earthquake in China), the author collected statistical data accordingly (e.g., on the first, second and third days after the earthquake) for use in future studies.

4.1.3 – Classification of the Top 50 Chinese/English News Reports

The above-mentioned top 50 news reports published by *the People's Daily Online* from the selected dates were clustered in accordance with the classification scheme discussed in Chapter 2 to address the first research question. The biggest advantage of such a scheme is that it was in line with the original English channel names of *the People's Daily Online*, which made the data collection and analysis work much easier and more efficient. More importantly, results from this thesis (especially answers to the first research question) will provide more useful feedback to the editors. During the "clustering" process, the author decided to add the following categories: **Olympics, Earthquakes, Taiwan, Tibet, Crime, Environment, Odd**. Besides reports on Chinese Politics or World Politics, all other categories of reports were not divided into "Chinese" and "other countries", due to the fact that the vast majority of the reports (more than 85% of the top 50 news) were on China. The merge of "Chinese" and "other countries" also applied to the top 50 query classification scheme.

To ensure the accuracy of the classification work by the author, several journalists from the *People's Daily Online* were invited to help group the Chinese and English news reports/queries in accordance with the revised classification scheme, which was explained in detail by the author to the second coder. The author classified the top 50 news reports/queries first, and then sent the original list without classification information to the second coder. After comparing the two classification results, the author noticed the inter-coder consistency rates were pretty

good (with an average of approximately 92.5%) for the data from the first 61 days (October and November 2007), and inconsistent coding was resolved quickly. The author then conducted the remaining classification duties alone.

4.1.4 – Classification of the Top 50 Chinese/English Queries

The WebTrends Log Analyzer provided the total numbers and contents of the queries submitted to the website search engine of the *People's Daily Online*, which is designed by the news site itself rather than the “powered by Google” appliance. This website search engine can execute keyword (single words or phrases) searches as well as Boolean searches (with “AND”, “OR”, and “BUT” operators). If no Boolean operator is explicitly specified, the space in between query terms will be interpreted as the Boolean operator “AND”. The meanings of the **phrases** forming each query were analyzed to identify the topics of the Chinese queries, while the meanings of **terms** were analyzed to identify the topics of the English ones. The author read the queries from the selected dates one by one, and then classified these queries by topic according to their meaning.

Considering the unique features of Chinese characters, the author could tell the meaning and topic of each query from the specific phrases it contained, which is different from the individual words or terms used in the English query analysis. For example, “腐败(corruption)” and “反腐败(anti corruption)” were grouped into the same category of “Chinese Politics”. All queries generated from the same report or group of news stories were clustered together. The relationships between the total number of these queries and the relevant top news items’ pageview numbers will be

discussed in Chapter 5.

It is possible that a query can be classified into more than one category (e.g., a vague query that just mentions the name of the Chinese President Hu Jintao). One solution to this problem is to exclude this type of vague query and record the number of omissions. Queries that clearly fell into two or more categories were assigned one of the main categories only if this category carried more than 50% of the weight (by the author's judgment and suggestions from reporters of the *People's Daily*), and omitted if they were evenly split between two categories. Preliminary results from analyzing data from the first two months (October and November 2007) showed that this type of query was few in number (on average about 2%, or one out of 50), and therefore not likely to significantly affect results.

Properly classifying the top queries, especially the "vague ones", requires us to locate "roots" of the top queries. The author adopted "retrospective lookup techniques" to deal with this issue: searching the previous three or even five days' news reports for proper clues. Once the searched keywords (queries) appeared in the top stories (full text rather than the title), we could easily group them into the same category of the "birthing" reports. If the "retrospective lookup techniques" still could not retrieve the related news reports or locate too many hits (as in the case of vague queries like "Beijing" or "Shanghai"), we placed them into the "other" category; as a result, no query was omitted from this study.

On average, approximately 20% of the top 50 Chinese queries were classified with the help of "retrospective lookup techniques". Following this analysis, there

were still several queries (approximately one out of the 50) remaining which had to be assigned to the "other" category. For the English queries, the above-mentioned measures were slightly higher, having values of 30% and 4% respectively. The "retrospective lookup techniques" might not have been the best method by which to classify the top 50 queries; however, they are perhaps the most effective way at the current stage for the author to find what a specific query was intended to search for.

It is safe to assume that online readers nowadays go to frequently updated news websites for the latest information, and their searching behaviors will be focused on the news-related topics rather than other issues. Such a phenomenon causes news website search engines to differ from general purpose search engines like Google or Yahoo!. A previous study published by Nature News (Ball, 2005) showed that the lifespan of online news reports is around two days. For the current thesis, more than 95% of the Chinese/English queries were related to current events, which proved the applicability of the "retrospective lookup techniques" adopted by the author.

4.1.5 – Choosing and Classifying Reports on the United States and Japan

In this thesis, "reports on the United States and Japan" not only included reports on events which happened within the geographic boundary of the two countries, but also covered news stories regarding U.S./Japanese citizens and interests abroad, such as the U.S./Japanese troops' movements toward other countries, especially in "hot spots" such as the Middle East, the Korean Peninsula, or near the Taiwan Strait.

In this thesis, the "negative news reports" on the United States and Japan were categorized into the following two groups:

- The first category includes reports on critiques/suspensions/worries about the U.S./Japanese governments and policies (especially international and economic ones), as well as accidents/disasters involving human-factors;
- The second category contains reports on natural disasters that took place in the United States and Japan, as well as those that happened outside these territories but with casualties of the two countries' citizens.

Considering that readers will generally pay much more attention to natural disasters (such as floods, earthquakes, or wild fires) around the world, and there were not enough such news items (around 10 for the United States and 6 for Japan) during the data collection period, the "negative news reports" in this thesis were limited to reports on the "events involving human or governmental errors", such as plane crashes, traffic accidents, crimes, political turmoil, financial crisis, and attacks against the U.S. soldiers/citizens in Iraq, Afghanistan or other nations.

Meanwhile, "non-negative reports on the United States/Japan" were those with positive or neutral stances, which include:

- Articles welcoming, supporting or introducing the U.S./Japanese government's domestic or international policies, activities, and movements;
- Reports focusing on the achievements of the U.S./Japanese military, scientific research, athletes, education, and cultures, etc;
- All other stories on events in the United States/Japan with a neutral stance; e.g., general elections;

4.1.5.1 – Collecting U.S./Japan-Related Stories

Given that the WebTrends Log Analyzer for this thesis only provides the top 50 news reports' pageview numbers each day, the author first tried to collect one negative and one non-negative United States/Japan-related Chinese report from the same day's top 50 list. Once there were no such paired stories available from the same day, the author retrieved the needed pair within the shortest time period, with the restriction that the interval between the two stories should be no more than 10 days to reduce the possibility of visit fluctuations. Under these circumstances, when there was more than one report in a particular category, only the one with the highest pageview number was chosen.

The sample sizes for the two categories of news reports were kept equal: 30 negative and 30 non-negative stories, so as to ensure the power of the statistical tests. Two correspondents with the *People's Daily* in Washington and Tokyo were invited to re-classify the collected negative and non-negative U.S./Japanese-related stories respectively, which had been previously classified by the author, with an aim of conducting an inter-coder consistency check and minimizing possible subjective factors. The inter-coder consistency was found to be very high, as there were differences in categorization for only two of the 60 United States-oriented reports and one of the 60 Japan-related news items. After discussions, the coders reached 100% agreement on the classification of the news reports.

However, the author could not collect the same amount of negative and non-negative U.S./Japanese news reports from the English edition of the *People's Daily*

Online in line with the previously mentioned criteria. Therefore, this thesis only investigated the Chinese readers' interests in United States/Japan-related issues. Please refer to Appendices 3-6 for the full list of U.S./Japanese-related news titles, which were translated into English by the author.

4.1.5.2 – Collecting Comments Posted to U.S./Japan-Related Reports

The variance of pageview numbers is an indirect measure of the Chinese readers' interests in the U.S./Japanese-related issues, as well as the public's mood. A more direct gauge of public sentiment would be the contents of comments posted to the news reports. This thesis further investigated readers' interests through a qualitative content analysis of the comments posted to the reports on the United States or Japan.

For each of the selected U.S./Japanese-related reports, the author randomly sampled four comments posted by the readers. Thus, there were 240 comments on the United States/Japan-related non-negative news reports and 240 comments on the United States/Japan-related negative news reports. Following scheme was adopted to classify the "opinions left by real readers":

- Positive Comments: those supporting the United States or Japan;
- Negative Comments: those with anti-American/Japanese sentiments;
- Neutral Comments: all others objectively analyzed United States/Japan-related issues, and were neutral in tone (neither positive nor negative).

One editor from the *People's Daily Online* helped the author classify the collected comments on the United States/Japan-related stories, which were then re-

classified by the author, with the aim of conducting an inter-coder consistency check and minimizing possible subjective factors. The inter-coder consistency was found to be high, as only four out of the 240 U.S.-oriented comments and three out of the 240 Japanese-related comments were classified inconsistently between the two coders. After brief discussions, the coders reached 100% agreement.

4.1.5.3 – Collecting Russia-Related News Reports and Comments

The *People's Daily Online* published plenty of reports on Russia, which offered us a unique contrast example to study the readers' interests in certain nations. During the data collection period of this thesis, one of the most relevant events was the Georgian War in August 2008. By analyzing the reactions of Chinese readers to news reports on this regional war, we could tell whether these people were really "anti-War" or just merely "anti-American", since the Chinese online readers showed strong opposition to the U.S.-led wars in Iraq and Afghanistan.

Collecting and classifying the Russia-related data (both news reports and comments) was straightforward: we just followed the rules set in the previous sections on the U.S./Japanese issues: 30 negative and 30 non-negative news reports on Russia-related events were collected, and each report also had four randomly sampled comments to be analyzed. The inter-coder consistency rates for the Russia-related reports and comments were also very high, which were 56 out of 60 for the news items and 233 out of 240 for the comments. The coders then reached 100% agreement after discussions. See Appendices 7-8 for Russia-related reports.

4.1.6 – Collecting Homepage Position Data for the Top News Reports

A total of 12 randomly selected days' worth of Web server log statistical data from the *People's Daily Online* were collected from the previously mentioned 129 dates between October 2007 and September 2008, which included the daily pageview numbers for the top 50 Chinese and English stories. With the help of Internet Explorer 7.0, the author manually saved the Chinese and English homepages four times on each of the 129 selected dates (9am, 12pm, 3pm and 21pm, Beijing Time), in accordance with the updating frequencies of the *People's Daily Online*, and then retrieved the position information for the top 50 news items on the homepages.

The *People's Daily Online's* Chinese homepage (from 2005 to 2008) was divided into the following sections: Homepage Pic, A section, B section, C section. The English homepage was also divided as well, into Homepage Pic, A section, B section and C section. For both of the Chinese and English homepages, the C sections are not completely visible above the fold. It is obvious that except for the Homepage Pic, all the other codings here have different meanings on the Chinese and English homepages (see Figures 4-1 and 4-2 on the next page for details). According to the editorial rules of the *People's Daily Online*, the Homepage Pic and A Sections of the Chinese and English homepages were "reserved" for the most important news stories. In the meantime, the C Sections on the English homepages were for the list of latest updated reports. The C Sections on the

Chinese homepages were “reserved” for articles on relatively “light” topics, such as health care or culture/life/society issues.



Figure 4-1: Homepage of the *People's Daily Online* (Chinese Edition)

There are two methods by which to study the influence of homepage positions on pageviews for the top news items: the first one pools all position data together and then runs proper statistical tests to compare the differences in the top 50 news reports' pageview numbers. It was easy and feasible for us to analyze the 12 days' worth of data in this way. However, if there are more data to be analyzed (e.g., 1000 days), such an analysis method would be ill-advised and less efficient.

Fortunately, we do have an alternative method to deal with such an issue. The collected top 50 news reports' homepage position data were not evenly distributed among the various groups; thus, we should always use the smallest number of report counts for each position as the sample size of that day. For example, if one

homepage position yields only one news item from the list of the top 50 reports (which is the smallest among various positions on the same day), we should also choose the top sample from the other positions' candidates. This is one of the most efficient methods to ensure adequate power of the results with the help of equal-sized samples. Altogether, 116 Chinese and 138 English reports were analyzed to answer the RQ3.

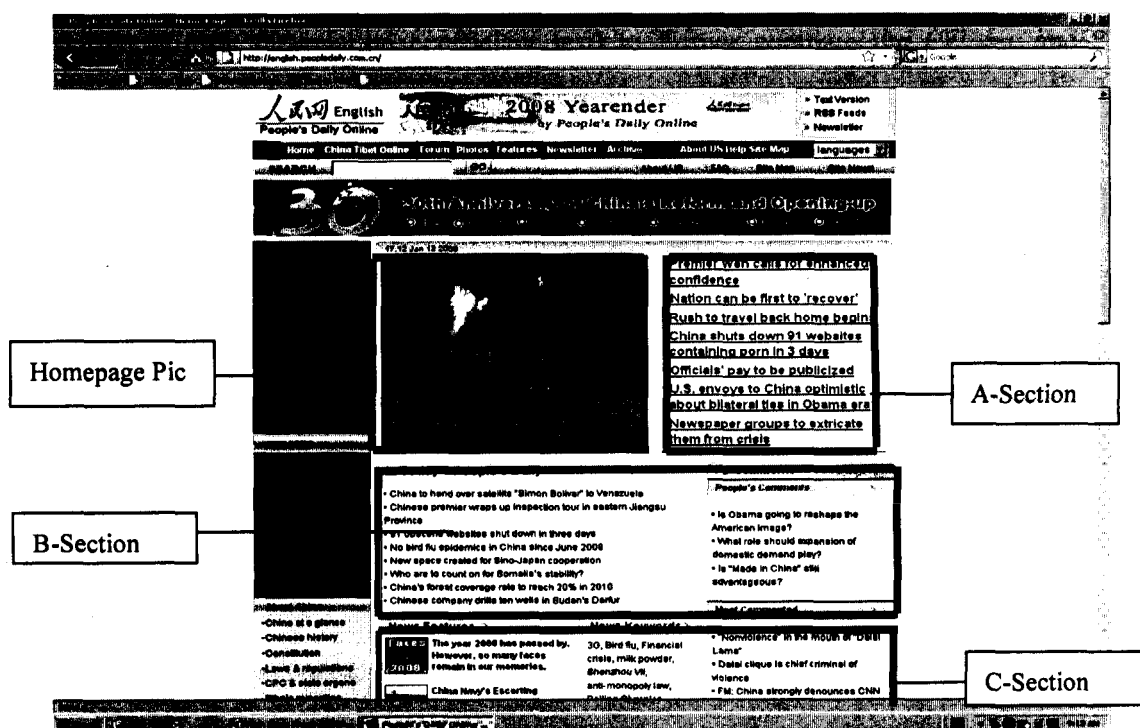


Figure 4-2: Homepage of the *People's Daily Online* (English Edition)

4.1.7 – Pairing the Chinese and English Reports on the Same Events

For each of the previously mentioned 12 days' Chinese and English top 50 reports, the author manually identified those on the same events, since the English news items of the *People's Daily Online* were translated from this online newspaper's Chinese ones. The pageview numbers of the selected titles were

collected and saved in an Excel file. The pageview numbers of the Chinese stories were listed in the first column and their English counterparts' pageview numbers in the second column. Thus the two numbers in each row were the pageview numbers for the same story's Chinese and English versions respectively. For each of the 12 randomly selected dates, the average number of same-story pairs was around 18 out of the top 50 Chinese/English news reports, which was sufficient for us to conduct statistical tests (correlation coefficient test) to further compare the Chinese and English readers' interests.

4.2 – Collecting and Classifying Online Survey Results

Since 2005, the *People's Daily Online* has been launching surveys prior to the annual National People's Congress (NPC) sessions held in March to learn the public opinion and expectations of the Chinese government. Designs of these surveys are straightforward: the readers can choose multiple options, and then submit their selections to the server. By clicking on the "show the results" button, everyone who visits the survey pages, no matter whether they voted or not, can view which category of option attracted more votes from the readers. In this way, the author easily collected two years' survey results (in 2008 and 2009) by the end of the day (23:59pm, Beijing Time) before the opening ceremony of the annual sessions, when the surveys were removed from the homepage of the *People's Daily Online*.

All the survey options were translated into English and then classified by the author in accordance with the previously discussed top news reports/queries classification scheme. After the initial classifications, all survey questions belonging

to the same category (e.g., Chinese Politics) were merged together. Both the 2008 and 2009 survey results were compared with the server log statistical data collected during the annual NPC session, thus enabling a deeper exploration of readers' interests by shedding light from different perspectives.

To remove "votes" to the surveys from crawlers or spiders, all survey participants had to first enter "eye-readable" numbers printed on a small picture, and then submit their votes to the server. There was no limitation on the number of votes one reader could cast: choosing the whole list or only one vote were both acceptable. However, there were some limitations on "repeating votes" (i.e., the same IP address with identical browser/hardware settings could only vote once every one hour), and the internal votes from the *People's Daily Online's* staff were not included in the displayed results, both in order to guarantee the quality of the data.

4.3 – Collecting and Classifying External Inlinks

There are two types of inlinks: total inlinks and external inlinks. Total inlinks include all links pointing to a particular site, while external inlinks include only links coming from domains outside of the site in question. In other words, external inlinks do not include links originated within a site being studied, such as the "back to home" type of navigational links (Gao & Vaughan, 2005). This study only examined external inlinks pointing to the *People's Daily Online* or other leading news websites.

4.3.1 – Choosing Appropriate Search Engines to Retrieve Inlinks

Two popular commercial search engines (Google and Yahoo!) could conduct inlink count searches during the period of data collection (from June to October 2008). However, Google's performance on external inlink searches was still not satisfactory (outnumbered significantly by Yahoo!); see Table 4-1 for details (data collected on June 18, 2008). AltaVista and AllTheWeb used to be employed for web link studies, but they were acquired by Yahoo! in early 2004, and their databases became a subset of Yahoo!. Hence, these two search engines were not considered for data collection for this thesis. Another popular search engine – MSN (Windows Live) – stopped offering the inlink search function in March 2007 (Seidman, 2007). Other leading Chinese search engines, such as Baidu (www.baidu.com) or Sohu (www.sohu.com), cannot perform any inlink-related searches at all. Thus, Yahoo! (both the China and Global versions) were chosen as the data collection tools for the external inlink study section of this thesis.

Table 4-1: Inlink Search Results of Yahoo.com and Google.com

Search Engine	Query to Retrieve Inlinks	Results
Yahoo!	link: http://www.people.com.cn - site: people.com.cn	1,780,000 (top 1000 results displayed)
Google	link: people.com.cn	59,700 (753 results displayed)

4.3.2 – Using Proper Queries to Retrieve Inlinks

Since the current study intended to collect as many external inlinks pointing to the *People's Daily Online's* Chinese and English reports as possible, the author employed both Yahoo!'s English (www.yahoo.com) and Chinese versions

(www.yahoo.cn) to test their performance in retrieving external inlinks to the *People's Daily Online*. Four rounds of tests were conducted from June 2008 to September 2008. Since the total inlinks retrieved by Yahoo (both Chinese and English versions) were fairly stable, the author decided to classify the first 1000 inlinks retrieved in September 2008.

The two target sites were the *People's Daily Online* Chinese Edition (<http://www.people.com.cn>), and the *People's Daily Online* English Edition (<http://english.people.com.cn>). The first query used to locate external inlinks through Yahoo! was **"link: <http://www.people.com.cn> -site:people.com.cn"**, which only retrieved external inlinks pointing to the homepage of the *People's Daily Online* Chinese Edition. Inlinks pointing to other pages within this domain were not included. However, another Yahoo! query: **"linkdomain:people.com.cn -site:people.com.cn"** retrieved all external inlinks pointing to any pages within this domain. Same queries were also submitted to Yahoo's Chinese site (www.yahoo.cn) and retrieved significantly different URLs. In addition, similar queries for retrieving the external inlinks to the English homepage and all pages within the domain were submitted to Yahoo's English and Chinese sites as well (see Table 4-2 for details).

Table 4-2: Inlink Search Results of Yahoo.com and Yahoo.cn

Search Engine	Search Queries			
	linkdomain:people.com.cn -site:people.com.cn		linkdomain:english.people.com.cn -site:people.com.cn	
	The Whole Web	Chinese Only	The Whole Web	Chinese Only
Yahoo.com	2,740,000	1,970,000	308,000	283,000
Yahoo.cn	5,420,000	N/A	24,400	N/A

From Table 4-2, we can infer that Yahoo.com should be applied to collect the inlinks to the English Edition of the *People's Daily Online*, while Yahoo.cn is a good choice to search inlinks to the Chinese pages, thus highlighting that different search engines index different websites and that there are advantages to using multiple search engines for data collection (Vaughan & Zhang, 2007). "N/A" in Table 4-2 indicates that Yahoo.cn cannot limit its results to a specific language, such as "Chinese" or "English".

Therefore, Yahoo's "linkdomain" command helped us collect inlinks pointing to the news reports as well as homepages of the *People's Daily Online*. However, for each round of search, Yahoo! (both the English and Chinese versions) could only list the first 1000 results, which might have led to some fairly uninformed conclusions. Even with the help of advanced search functions (country and language options) to narrow down the results, the number of retrieved inlinks was still far more than the maximum 1000 URLs displayed by Yahoo.com. As there was no similar advanced option for the Yahoo.cn, we had to use the original list of 1000 to conduct our study.

On the other hand, no search engine can index the whole Web; therefore, what we had was only a small sample made available by Yahoo!. Despite this, the 1000 inlinks retrieved by Yahoo!, though they might not have reflected the real distribution of the inlinks, still offered us some valuable information on readers' interests from other aspects. However, conclusions drawn from such publicly accessible data might be limited.

Considering Yahoo! (both English and Chinese versions) displayed the first 1000 hits of each search, the author randomly chose one from every five results to study the inlinks' originating countries, types of host pages, types of linked news, and the purposes of creating the inlinks. A similar analysis method was previously adopted by Gao and Vaughan (2005).

4.3.3 – Classifying the Retrieved Inlinks

Retrieved URLs were merged into two sets of data: those with external inlinks to the *People's Daily Online* Chinese Edition, and those with external inlinks to the *People's Daily Online* English Edition. Since this thesis only sampled 200 Chinese and 200 English URLs retrieved by Yahoo! (English and Chinese versions), the author manually examined the retrieved URLs of these linking pages (pages that initiated the inlinks), which were copied into an Excel file for further data filtering as follows.

All sampled URLs returned by Yahoo! search were analyzed, after excluding dead or outdated inlinks (i.e., the retrieved page originally had an external inlink to the *People's Daily Online*, but the content of the page had been changed since Yahoo! indexed it and thus current page had no inlink to the site under study). All duplicate URLs in the merged lists of Yahoo! search results were identified using Excel's "find" function and subsequently removed. Similar techniques were previously adopted by Gao and Vaughan (2005).

To establish online profiles of the *People's Daily Online* (Chinese and English editions), each linking page was classified according to the following attributes:

- (1) Language: e.g., Chinese, English, Japanese, French, etc;
- (2) Country: e.g., China, U.S., U.K., France, Japan, Russia, etc;
- (3) Type of site: e.g., corporate site, government site, etc., see Appendix 9;
- (4) Category of news or information inlinked;
- (5) Reasons for/Purposes of linking.

In the above classification scheme, "country" was determined by the location of the individual or organization that was responsible for the content of the page, rather than by the physical location of the web server on which the page was stored. This is consistent with the OCLC (2002) method of country classification. Such information was retrieved from the "about us" or "contact us" sections. For the pages without country information, "Unknown" was assigned as their place of origin. "Type of site" was determined by the content of the site, rather than by its domain name (e.g., .com, .gov, .org or .edu).

The category of inlinked news or information adopted the same classification scheme as the top 50 news reports with the highest daily pageview numbers. The "inlinked" content from the *People's Daily Online* offered us the opportunity to learn what kind of information was more likely to be "cited" by other websites.

The reasons for/purposes of linking include the following:

- To list news media;
- To indicate the source of news reports;
- To provide further information;
- To show partnership.

If there was no obvious indication of the inlinks to the *People's Daily Online* (such as the name of the news sites on the retrieved webpages), the author let the browser open the "source" of the webpages, and then try to locate "people.com.cn" to retrieve the desired URLs with the help of the "find" function. For example, some websites only used "other resources" as the name of inlinks to the *People's Daily Online*, or used the title/keywords of the cited news reports as the text of the hyperlink, which did not tell us the URLs of the inlinked page directly.

4.3.4 – Collecting Interlinking Data among News Websites

There are some other interesting and useful queries to explore inlinks to the *People's Daily Online* from some specific domains. For example, we could retrieve external inlinks to the *People's Daily Online* from the *New York Times* on the Web, one of the most influential online newspapers in the United States, or vice versa. We could use the following query to establish the interlinking structure between the *People's Daily Online* and its Chinese partners/overseas counterparts on the Web: linkdomain:people.com.cn site:abc.com. This query not only worked for Yahoo.com, but was also supported by Yahoo.cn (the Chinese Yahoo!) during the last stage of data collection (August 2009). From such data, we were able to find some clues to study the "interlinking" or the "link journalism" phenomenon between the leading online newspapers in the two countries with the most netizens: China and the United States.

At the bottom of the *People's Daily Online's* (Chinese Edition) homepage, there are dozens of online media agencies listed within the "links to

friends/partners" section; therefore, the inter-link counts retrieved by the Chinese version of Yahoo! (www.yahoo.cn) could be employed to test the "friendship" among these so-called partners, since most of them are in fact the competitors of the *People's Daily Online*.

4.4 – Collecting Online Reprint Data

Online reprint data here refers to how many times a top news story released by the *People's Daily Online* (Chinese or English edition) was possibly republished by other websites around the world. Compared with the "cited contents" indicated by external inlinks, the online reprints are the "full text versions" of the original news stories, and may or may not bear external inlinks to the website being examined.

Google.com and the Chinese search engine Baidu.com were the two candidates at the initial stage of this study to retrieve online reprints rates of the top 50 English/Chinese news reports, since they are the most widely recognized search engines to find English and Chinese webpages. After four rounds of data collection and testing in November 2007, Google.com always returned more online reprints than Baidu.com (for both of the Chinese and English news stories, at least 20 percent more); thus, Google.com was eventually chosen in this study to retrieve the top 50 Chinese/English news reports' online reprint rates.

The queries for the reprints were the "titles in quotation marks" and the whole Web was searched for more complete results. Titles here refer to those titles of the top 50 Chinese and English news reports from the *People's Daily Online*. The author tried the Chinese and English versions of Google for the online reprint data

of both English and Chinese reports, and got the same results. Hence, there was no need to switch search engines for the English and Chinese titles of the top 50 reports. The author noticed that the online reprint rates generally reached their peak after two days, and thus this period was chosen for data collection for the top 50 news reports' online reprint rates.

Chapter 5 – Data Analysis and Results

5.1 – Analyzing Web Server Log Statistics

After collecting all the required data, the author first tested the skewness of the server log statistics (with the help of the data set's histogram), so as to decide whether to employ parametric or the non-parametric statistical tests to process the data, and then addressed the five research questions of this thesis. SPSS (the 16th version) was the software used to run all statistical tests.

5.1.1 – Learning Readers' General Interests from Server Logs (RQ1)

According to Vaughan (2001, p.23), a histogram is “a bar graph for the frequency distribution of a group of data” (e.g., Figure 5-1 and Figure 5-2), and “a symmetrical distribution typically peaks around the middle of the histogram while a skewed distribution peaks at the right or left end”. The shape of the histogram (skewed or symmetrical) drawn by the SPSS was a factor in deciding the type of statistical analysis (parametric or nonparametric) used for this doctoral thesis.

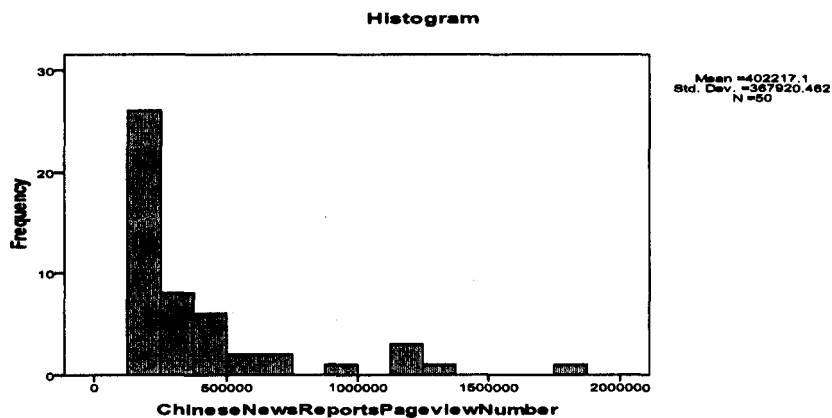


Figure 5-1: Histogram for the Top 50 Chinese News Reports' Pageview Numbers

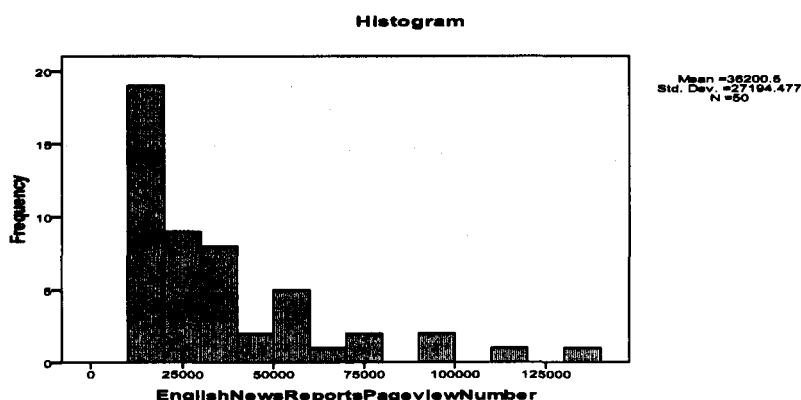


Figure 5-2: Histogram for the Top 50 English News Reports' Pageview Numbers

After classifying all the top 50 English/Chinese news stories and queries of the 129 days' server log statistical data sets, SPSS was employed to do frequencies tests for these groups. The histograms for the top 50 Chinese/English news reports and queries (see Figure 5-1 to Figure 5-4 for details) tell us that the frequency distributions of these data sets were highly skewed. Therefore, proper nonparametric statistical tests were used to examine the data sets reported in this chapter.

5.1.1.1 – Distribution Patterns of Top 50 Reports and Queries

We first tested the frequency distributions of the top 50 Chinese/English news reports measured by the number of stories within each news report's topic category. We also conducted the same test on the top 50 news reports within each of the same topic categories as measured by pageview numbers. Statistical correlation tests were used to examine the relationships between these two data sets. Statistical correlations here and in the following sections of this chapter all address issues of the strength, the type, and the statistical significance of the relationships between two variables (Vaughan, 2001).

Considering the highly skewed pageview numbers for the top news stories, the non-parametric Spearman's Rho test was used, and revealed very strong positive relationships between the two data sets (all correlation coefficient values were very close to 1). A similar pattern also existed in the frequency distributions of the Chinese/English top 50 queries (see Table 5-1 for details).

Table 5-1: Spearman's Rho Test Results

	Top 50 Reports		Top 50 Queries	
	Correlation coefficient	P-Value	Correlation coefficient	P-Value
Chinese	0.991	<0.001	0.983	<0.001
English	0.985	<0.001	0.967	<0.001

Thus, it was safe for us to use distribution patterns measured by the number of reports/queries in each category to investigate the general interests of the readers of the *People's Daily Online*. Compared with the pageview number of each news report's category, the number of reports from each category is a publicly available data source for us to access from the leading news websites, given that few online news agencies want to share the pageview numbers of their reports with the public. These statistically significant results helped us to find an alternative data set for the server log statistics: the most popular news reports list, which perfectly reflected the distribution patterns of readers' interests measured by pageview numbers (almost same ranking).

5.1.1.2 – Examining the Top 50 Chinese/English Reports

From Table 5-2, we learned that Chinese Politics, followed by Business and Olympics, were the most popular topics for the Chinese readers among the 6450

sampled reports, which was in line with the *People's Daily Online's* "governmental mouth piece" status in China. Such a pattern also mirrored the Chinese people's interest in the "hot events" from 2007 to 2008, such as the changing of China's leadership, the severe decline of the Chinese stock market, the rapid rise in housing and oil/gas prices, as well as the continuous deterioration of the U.S. financial market.

It is worth noting that the Beijing Olympic Games received more (possibly even overwhelming) coverage in August 2008. The *People's Daily Online* opened special columns on its homepages to report the Olympic news. Another consideration was the earthquake issue; we noticed that the intensive reports on the Sichuan Earthquake-related issues lasted for more than one month. These two "long-term hot events" contributed many more visits to the *People's Daily Online* than other issues, and therefore were treated as "atypical data". Since data was collected for the whole month of August 2008, we removed it to see what would happen in the results.

For the Chinese readers, Sci-Tech-Edu issues got a lot of attention due to the successful launching of the Chinese lunar orbiter "Chang'er" in 2007 as well as the first spacewalk by Chinese astronauts from "Shenzhou-7" within one year. Meanwhile, some educational issues, such as the national university/college entrance exam, also directed visitors to the *People's Daily Online* for more authoritative information.

Table 5-2: Frequency Distribution of Top 50 Chinese Reports from 129 Days

News Report Category	Frequency	Percentage
Chinese Politics	1579	24.5%
Business	751	11.6%
Olympics	713	11.1%
Taiwan	657	10.2%
World Politics	511	7.9%
Culture/Life/Society	431	6.7%
Entertainment	393	6.1%
Sci-Tech-Edu	346	5.3%
Earthquake	213	3.3%
Health	182	2.8%
Accident/Disaster	150	2.3%
Tibet	112	1.7%
Environment	108	1.7%
Odd	96	1.5%
Crime	75	1.2%
Other	70	1.1%
Sports	64	1%
Total	6450	100%

In 2008, some natural disasters and human-factor involved accidents other than the Sichuan Earthquake happened across China, which raised the ranking of such news reports published by the *People's Daily Online*. Please note that Sports in Table 5-2 referred to stories of athletic events other than those of the Beijing Olympic Games, which told us that the *People's Daily Online* was not an ideal medium for Chinese readers to receive sports news during ordinary days.

Besides the Olympics as well as the Business news reports, it was very interesting to point out that the English readers tended to read relatively more World Politics news reports but less Chinese Politics stories than their Chinese

counterparts (see Table 5-2 and Table 5-3 for details), since the World Politics news reports ranked first and Chinese Politics news reports ranked fourth in Table 5-3.

One possible explanation for this issue is that the Chinese government routinely blocked many foreign news websites in China, such as the *BBC*, the *New York Times* or *CNN*. The author personally could not visit such sites without employing proxy servers in Beijing, and the Chinese government did not completely deny such accusations (BBC, 2008). Thus, the English readers in China might have had to access the *People's Daily Online* for political news reports on other countries.

Such an interesting pattern of results might also suggest that the English edition of the *People's Daily Online* does serve some of its purposes in expanding its readers' base. English readers in China might be more likely to visit news reports about their motherland than stories on China's domestic political events. Such an explanation, however, is just the author's personal speculation. For example, if I found a title mentioning "China" while browsing the *New York Times'* website, I would very likely click on it before reading other news reports. Therefore, more statistical tests (i.e., triangulation of survey results and server log statistical data of the English Edition) are needed in the future to reach a more decisive conclusion on this issue.

Table 5-3: Frequency Distribution of Top 50 English News Stories from 129 Days

News Report Category	Frequency	Percentage
World Politics	1372	21.3%
Business	935	14.5%
Olympics	899	13.9%
Chinese Politics	732	11.3%
Sci-Tech-Edu	401	6.2%
Accident/Disaster	295	4.6%
Life/Culture/Society	276	4.3%
Health	241	3.7%
Tibet	229	3.6%
Entertainment	218	3.4%
Sports	210	3.3%
Environment	182	2.8%
Earthquake	165	2.6%
Other	114	1.8%
Odd	85	1.3%
Crime	49	0.8%
Taiwan	47	0.7%
Total	6450	100%

In terms of the reports related to Taiwan and Tibet, there were two major events that happened involving these regions: the changing of leadership in Taiwan and the protests over Tibet after the Lhasa riots during the Olympics Torch Relays. However, something noteworthy here was the difference in readers' interests in Taiwan and Tibet issues. It was obvious that as a whole, the Chinese news readers cared about Taiwan (10.2%) more than Tibet (1.7%), and the English readers were opposite (3.6% on Tibet vs. 0.7% on Taiwan), from October 2007 to September 2008. This offers us insight into why there were so many protests during the Olympics Torch Relays around the world.

While this is speculation, it is possible that such a pattern of results is reflective of the reader's level of understanding of these issues: some English readers do not really understand the subtle relationships between the Chinese Mainland and Taiwan; many of the English readers think Taiwan is an independent state rather than "an in-alienable part of China", as claimed by the average Chinese readers. Since most of these English readers are not from Taiwan, they do not care about such "world political issues", which is in line with the previous finding on the English readers' interests in World Politics.

Contrary to this, most of the average Chinese readers hope Taiwan will reunite with the Chinese mainland some day, and do not believe Tibet should be granted any kind of independence or the highly autonomous status enjoyed by Hong Kong. However, the author found that many English readers believe Tibet has been an "occupied state" since 1951, which should be granted full independence. This difference of opinion might explain why the English readers accessed relatively more news reports on Tibet, and the Chinese readers accessed more news reports on Taiwan.

Another likely explanation for the "popularity difference" in Taiwan/Tibet-related reports between Chinese and English readers is that the Chinese readers might have cared more about "WHY" something happened or the implications of certain news events, whereas the English readers might have cared more about "WHAT" happened. This could explain why "reports on Tibet" ranked higher, since these reports published by the *People's Daily Online* (in both Chinese and English)

were more about “WHAT” happened without detailed elaborations of their implications. However, formal statistical tests (i.e., triangulation of survey results and server log statistical data of the English Edition) are needed in the future to reach a more decisive conclusion on this issue.

The ranking for the English Sports news reports was relatively higher than the ranking for Chinese articles of this category, which may be due to the following reasons: the *People's Daily Online* is a political and business issues-oriented news site, which is not an ideal Chinese medium for such a “light” topic. However, there were only about a dozen big English news sites in China, far less than their Chinese counterparts, and the access speed to the overseas news site was rather slow due to the “Great Fire Walls” set by the Chinese government to filter “information of no good to the stability of the country or a healthy society” (BBC, 2008). Thus, the English readers had to rely on the *People's Daily Online* to receive Sports news reports.

After removing the August 2008 data, the author found the majority of the top rankings for the Chinese news reports unchanged: only the “Olympics” and “Crime” categories went down to the end of the list. As there were some serious criminal events that happened across China in August 2009 that received intensive reportage by the *People's Daily Online*, these were “deleted” along with the overwhelming number of news items concerning the Olympics (see Table 5-4 for details).

Table 5-4: Distribution of Top 50 Chinese Reports excluding August 2008

News Report Category	Frequency	Percentage
Chinese Politics	1293	26.4%
Business	636	13%
Taiwan	570	11.6%
World Politics	391	8.0%
Culture/Life/Society	358	7.3%
Entertainment	354	7.2%
Sci-Tech-Edu	309	6.3%
Earthquake	199	4.1%
Health	150	3.1%
Accident/Disaster	120	2.4%
Tibet	110	2.2%
Environment	98	2%
Odd	77	1.6%
Other	51	1.0%
Sports	64	1.3%
Crime	61	1.2%
Olympics	59	1.2%
Total	4900	100%

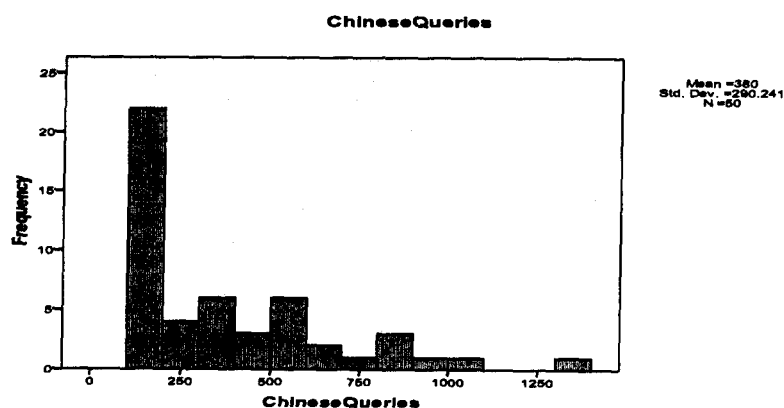
A similar pattern of results was also observed for the English top news reports' rankings. Such results told us that the overall interests of the Chinese and English readers did not change significantly during the one-year data collection period, and the data from August 2008 just raised the ranks of the Olympics news reports to a much higher position. Since the 2008 Olympic Games was the first Olympics held in China, it was very natural for them to attract more attention from the readers (see Table 5-5 for details).

Table 5-5: Distribution of Top 50 English Reports excluding August 2008

News Report Category	Frequency	Percentage
World Politics	997	20.3%
Business	826	16.9%
Chinese Politics	666	13.6%
Sci-Tech-Edu	407	8.3%
Health	259	5.3%
Tibet	255	5.2%
Life/Culture/Society	246	5.0%
Accident/Disaster	235	4.8%
Sports	210	4.3%
Entertainment	185	3.8%
Environment	172	3.5%
Earthquake	164	3.3%
Olympics	101	2.1%
Odd	76	1.6%
Taiwan	46	0.9%
Crime	30	0.6%
Other	25	0.5%
Total	4900	100%

5.1.1.3 – Examining the Top 50 Chinese/English Queries

Considering the highly skewed data of the pageview numbers and query counts (see Figure 5-1 to Figure 5-4 for the histograms), the Spearman's Rho correlation test was employed to examine the relationships between the two variables.

**Figure 5-3: Histogram for the Top 50 Chinese Queries' Counts**

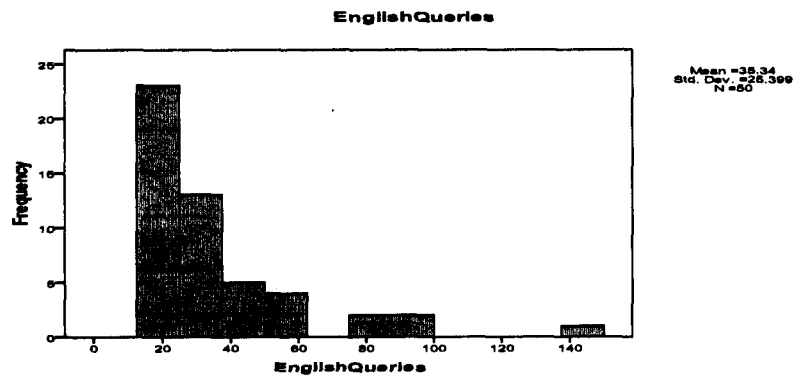


Figure 5-4: Histogram for the Top 50 English Queries' Counts

We found that the query counts were positively correlated ($P < 0.05$) with the pageviews of the top news reports from which they "originated" (**Correlation Coefficients:** 0.83 (Chinese) and 0.86 (English)). Since the two values were larger than 0.5 and closer to 1, there is a relatively strong relationship between the top news reports and the most searched queries (see Table 5-6 and Table 5-7 for details).

Table 5-6: Distribution of Top 50 Chinese Queries from 129 Days

Query Category	Frequency	Percentage
Chinese Politics	1465	22.7%
Business	1187	18.4%
Olympics	768	11.9%
Taiwan	621	9.6%
World Politics	450	7.0%
Entertainment	357	5.5%
Culture/Life/Society	284	4.4%
Sci-Tech-Edu	232	3.6%
Earthquake	229	3.6%
Tibet	164	2.5%
Accident/Disaster	132	2.0%
Other	129	2.0%
Health	113	1.8%
Environment	101	1.6%
Crime	97	1.5%
Sports	68	1.1%
Odd	53	0.8%
Total	6450	100%

Table 5-7: Distribution of Top 50 English Queries from 129 Days

Query Category	Frequency	Percentage
World Politics	1295	20.10%
Business	951	14.70%
Olympics	866	13.40%
Chinese Politics	590	9.10%
Earthquake	371	5.80%
Sci-Tech-Edu	367	5.70%
Accident/Disaster	320	5.00%
Tibet	269	4.20%
Other	265	4.10%
Culture/Life/Society	229	3.60%
Sports	193	3.00%
Entertainment	181	2.80%
Health	166	2.60%
Odd	147	2.30%
Environment	126	1.90%
Taiwan	59	0.90%
Crime	55	0.80%
Total	6450	100%

In other words, we could say that the queries from the news site search engine did reflect readers' interests and the influence of the top news reports. Although a larger pageview number for some categories of news items may or may not be the reason for a larger number of queries regarding similar issues, a significant correlation here did allow us to make some predictions about the queries based on the top news reports in future studies.

5.1.1.4 – Top News Reports' Pageview Numbers vs. Comment Counts

There were significant positive correlations ($P < 0.01$, Spearman's Rho test; see Table 5-8 for details) between the pageview number of each of the top 50 news reports and the counts of comments the same news report attracted as well as its

emailed frequencies. Thus, it is safe for us to say, at least for the *People's Daily Online*, that comment counts are good alternative data for us to study readers' interests. A significant correlation here also allowed us to make some predictions about the pageview numbers of top news reports based on their comment counts in future studies.

Table 5-8: Correlation between Pageviews and Comments/Emailed Frequencies

	Correlation Coefficients		P-Value
	Chinese	English	
Pageview and Comment Counts	0.692	0.757	<0.01
Pageview and Emailed Frequencies	0.625	0.689	<0.01

In a later section of this chapter, the author will conduct a content analysis of the comments on the sampled United States/Japan/Russia-related news stories, in order to explore the public attitude on these nations. Analyzing readers' comments is also a good way for the Chinese government to collect the people's opinion of some sensitive issues in China, such as the tainted baby formula scandals of September 2008 and the new Chinese leadership's overall performance.

5.1.2 – Chinese Readers' Interests in U.S./Japan-Related Events (RQ2)

Answering the second research question required us to find if there was a statistically significant difference between the two groups of variables ("Negative News Reports on the United States/Japan/Russia" vs. "Non-Negative News Reports on the United States/Japan/Russia") in terms of the pageview numbers these news items attracted during the 24-hour period. Thus, the "independent t-test" or the non-

parametric Mann-Whitney test was the appropriate one because the two samples of data were independent of each other.

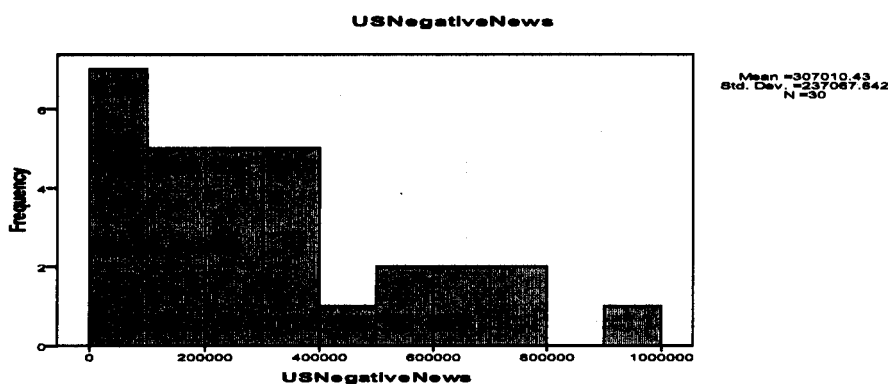


Figure 5-5: Histogram for U.S.-Related Negative News Reports' Pageview Numbers

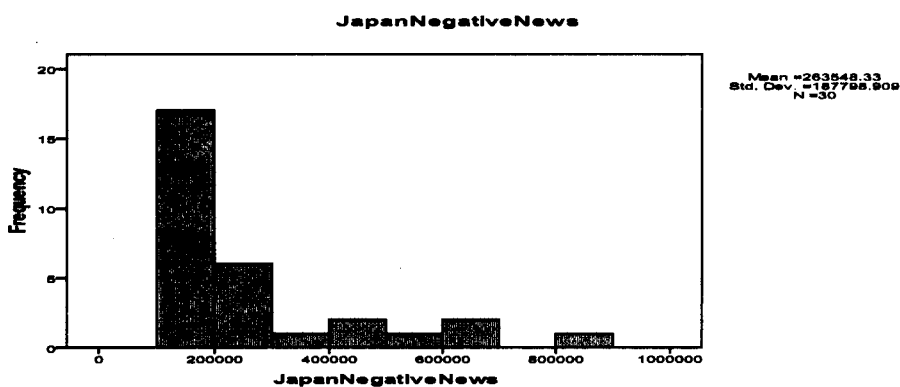


Figure 5-6: Histogram for Japan-Related Negative News Reports' Pageview Numbers

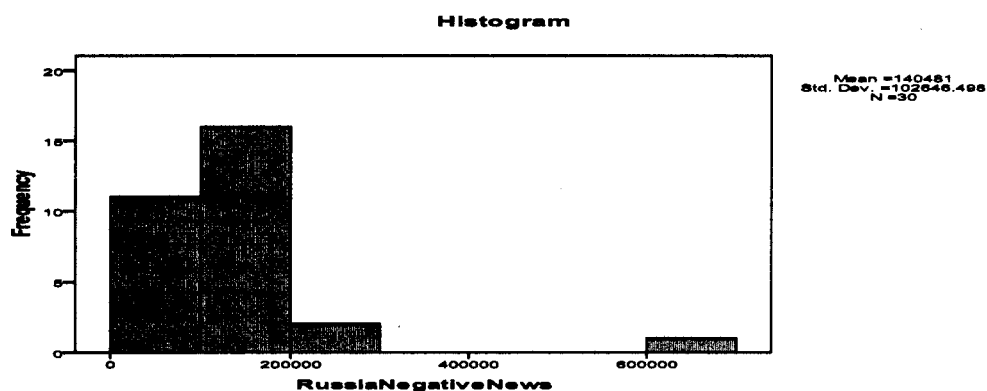


Figure 5-7: Histogram for Russia-Related Negative News Reports' Pageview Numbers

The frequency distributions of the pageview numbers of the United States/Japan/Russia-related stories were highly skewed (see Figure 5-5, Figure 5-6 and Figure 5-7 for details); therefore, the non-parametric Mann-Whitney test was applied to explore the differences between the pageview numbers of 30 Negative News reports and 30 Non-negative News items concerning United States/Japan/Russia-oriented events, and the P values for the three rounds of tests were all less than 0.01 (see Table 5-9, Table 5-10 and Table 5-11 for details).

Table 5-9: Mann-Whitney Test Results for U.S.-Related News Reports

News Report Type	Sample Size	Median of Pageviews	Mean	P-Value
Non-Negative News Reports of U.S. Issues	30	87140	199998.06	<0.001
Negative News Reports of U.S. Issues	30	221598	307010.43	

Table 5-10: Mann-Whitney Test Results for Japan-Related News Reports

News Report Type	Sample Size	Median of Pageviews	Mean	P-Value
Non-Negative News Reports of Japanese Issues	30	116493.5	159231.8	0.001
Negative News Reports of Japanese Issues	30	175833	283548.33	

Table 5-11: Mann-Whitney Test Results for Russia-Related News Reports

News Report Type	Sample Size	Median of Pageviews	Mean	P-Value
Non-Negative News Reports of Russian Issues	30	188742.00	270776.7	0.005
Negative News Reports of Russian Issues	30	102028.00	140481	

Thus, we can conclude that the pageview number differences between the Non-Negative news reports and the Negative ones on United States/Japan-related issues were statistically significant. As the medians of the pageview numbers for the negative stories were much higher than that of the non-negative ones, it is safe for us to note that at least for the *People's Daily Online* (Chinese Edition), Negative News on United States/Japan-oriented events tended to attract more visits than their Non-Negative counterparts.

In line with this trend, the pageview number differences between the non-negative news reports and the negative ones on Russia-related issues were also statistically significant, but tell us another story. The median pageview count for the non-negative items is much higher than that of the negatives items. It is safe for us to note that at least the Chinese news readers accessed more non-negative news reports on Russia-oriented events than they accessed their negative counterparts.

Comments posted to United States/Japan/Russia-related news reports offered us more important clues to study the Chinese readers' interests in the three countries (see Table 5-12 to Table 5-14 for details), and further supported results from the previous statistical tests reported on p.86 of this thesis: there is significant correlation between the top news reports' pageview numbers and comment counts.

Table 5-12: Comments Posted to U.S.-Related News Reports

Type of Report	Type of Comment			Total
	Positive	Negative	Neutral	
Negative News	0	109 (90.8%)	11 (9.2%)	120
Non-Negative News	32(26.7%)	45 (37.5%)	43(35.8%)	120
Total	32 (13.3%)	154 (64.2%)	54 (22.5%)	240

Table 5-13: Comments Posted to Japan-Related News Reports

Type of Report	Type of Comment			Total
	Positive	Negative	Neutral	
Negative News	16 (13.3%)	102 (85%)	2(1.7%)	120
Non-Negative News	42(35%)	36(30%)	42(35%)	120
Total	58 (24.2%)	138 (57.5%)	44 (18.3%)	240

Table 5-14: Comments Posted to Russia-Related News Reports

Type of Report	Type of Comment			Total
	Positive	Negative	Neutral	
Negative News	73(60.8%)	23(19.2%)	24(20%)	120
Non-Negative News	69(57.5%)	28(23%)	23(19.2%)	120
Total	142 (59.2%)	51(21.3%)	47 (19.5%)	240

A statistically significant relationship was found between type of comment and type of United States/Japan-related news story (Chi-Square Test, $p < 0.01$). However, we did not find a statistically significant relationship between type of comment and type of Russia-related news story (Chi-Square Test, $p=0.732$). It should be noted that the Chi-Square tests used here and in the coming sections of this chapter helped us examine the relationships between two variables with nominal or ordinal data (Vaughan, 2001).

It is obvious that for news reports on the United States, the Chinese readers posted their least positive remarks on both of the Negative (even a Zero!) and Non-Negative reports (only 26.7%), while Japan-related news stories got similar reactions from the Chinese readers (13.3% positive reactions for the Negative and 35% for the Non-Negative reports). For Russia, the Chinese readers left their most positive comments on both the Negative (60.8%) and Non-Negative events (57.5%). For the United States and Japan, readers' negative comments on the Negative news

items accounted for the large majority (90.8% for the United States and 85% for the Japanese ones). Overall, among the 240 randomly sampled comments posted to the U.S. news reports, 64.2% were negative. For the Japan-related stories, 57.5% of the 240 sampled comments were negative. However, 59.2% of the 240 comments posted to the Russia-oriented reports were positive. From these results, we can safely conclude that the Chinese readers tended to visit more negative reports than non-negative reports and left more negative comments than positive ones on United States/Japan-related issues. Although Japan sent much disaster-relief aid to China after the Sichuan Earthquake, the overall anti-Japanese sentiment among the Chinese readers were unchanged. For the Russia-oriented news reports, the Chinese readers expressed different attitudes: they accessed more non-negative news stories and left more positive comments on all of the reports.

It is worth noting that during the Georgian War in August 2008, comments posted by the Chinese readers overwhelmingly supported Russia's military actions against neighboring Georgia, and showed strong sympathies to Russia's loss of aircraft and soldiers. However, in the event of U.S. soldiers being killed in Iraq or Afghanistan, the Chinese readers always posted negative comments on such reports throughout the data collection period. These findings reflect the overall sentiments of the Chinese people toward these countries as described earlier in Chapter 2 (p.17) of this doctoral thesis (i.e., the Chinese netizens are anti-American rather than anti-War). Addressing Research Question Two further demonstrates the usefulness of Web server log data.

5.1.3 – Top News Reports' Pageviews and their Homepage Positions (RQ3)

Answering the third research question required us to find if the top news reports' various positions on the homepage, which are independent of each other, generated significantly different pageview numbers. Therefore, it was appropriate that either a one-way ANOVA or its non-parametric counterpart (the Kruskal-Wallis test) be used. For this statistical test, the independent variable was the top news reports' homepage positions, while the dependent variable was the number of pageviews generated by these news reports.

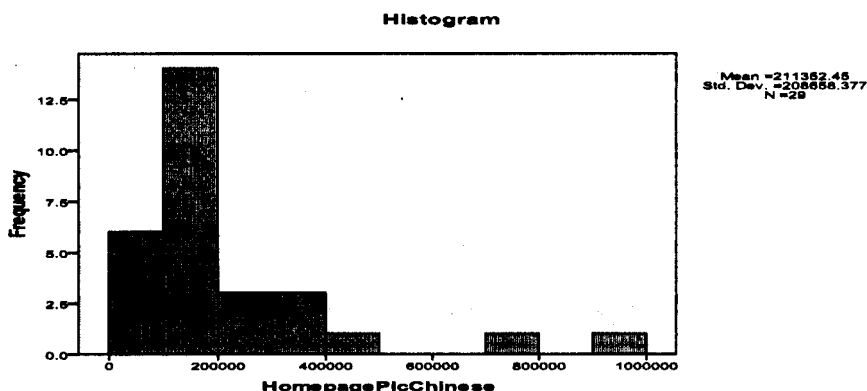


Figure 5-8: Histogram for Pageview Numbers for Chinese Reports on HomepagePic

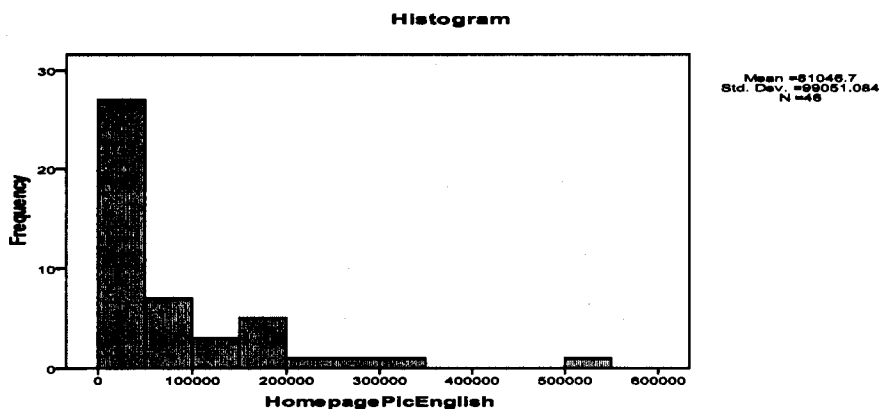


Figure 5-9: Histogram for Pageview Numbers for English Reports on Homepage Pic

Since the frequency distributions of the pageview numbers of the sampled top 50 Chinese/English news reports placed in various homepage positions were highly skewed (see Figure 5-8 and Figure 5-9 for details), the Kruskal-Wallis test was used to study the differences in pageview numbers across the news reports' various positions. This test was carried out for the Chinese and English editions of the *People's Daily Online*, and revealed some interesting findings.

For the Chinese homepages test, the P-value was less than 0.001, which told us that there were some significant differences among the pageview numbers of the four groups of news reports being compared. Although the test results did not specify the pattern of differences, judging from the median and mean of the pageview numbers of the four groups (see the third and fourth columns of Table 5-15), we still found useful pattern of differences.

Table 5-15: Kruskal-Wallis Test Results for Chinese Top Stories on Homepages

Homepage Position of Top Chinese Reports	Sample Size	Median of Pageview Numbers	Mean	P-Value
Homepage Pic	29	138470	211352.45	P<0.001
A	29	257629	339259.62	
B	29	531942	652622.17	
C	29	299251	363065.65	
Total	116			

It is safe to say that news items put on the "B" sections of the Chinese homepages attracted more visits than any other position because their median pageview number was much higher than the median of the other groups. However, there did not seem to be a significant difference between the A sections and C sections due to their fairly closed medians of pageview counts. Since the median of

pageview numbers attracted by titles placed in the Homepage Pic section seems to be lower than the same measure for the other sections, we concluded that for the Chinese Edition of the *People's Daily Online*, titles placed in the "least visible locations" of the homepage attracted more readership than the Homepage Pic and A sections, which the editors considered more important. It is clear that for the Chinese edition of the *People's Daily Online*, the editors' choices of "important" news reports **did not** match the readers' interests during the data collection period.

There was a different story for the English homepages. According to Table 5-16, we found that there did not seem to be a significant difference among the visits to top news items placed in various positions on the English homepages, although the P value of 0.122 is worth noting. The means and the medians of the pageviews numbers of the three groups of news reports placed in various homepage positions were relatively close.

Table 5-16: Kruskal-Wallis Test Results for English Top Reports on Homepages

Homepage Position of Top English Reports	Sample Size	Median of Pageview Numbers	Mean	P-Value
Homepage Pic	46	34885	81046.70	0.122
A	46	68987	105480.72	
C	46	48623	72109.93	
Total	138			

Results of the two Kruskal-Wallis tests showed that the Homepage Pic sections of the English and Chinese editions of the *People's Daily Online* did not attract more visits than plain text titles from other parts of the homepages.

It is obvious that news titles placed in the relatively less visible B section of the Chinese homepages and the A section of the English homepages both generated higher pageview numbers than news titles in the other positions. In addition, the Homepage Pic (photo) section of the Chinese and English homepages necessitates some revisions to play an even larger role, because it might be the case that some readers did not want to click the larger photo after viewing the miniature ones on the homepage.

For the Chinese Edition, the A section (the title with the largest font size) did not attract more visits. However, there is little room for the *People's Daily Online* to change this feature due to the government's regulations, which stipulate that only "the most important news reports in the editor's mind" can be placed in this position. However, the performance of the larger-font titles on the English homepage (A section) was much better, though news titles placed there were also stories of political or economic significance. Furthermore, as it was observed that news titles in the B Section of the English homepage never showed up in the top 50 news list, it is imperative that this section undergo major revisions to improve its performance.

5.1.4 – Chinese and English Readers' Interests (RQ4)

There were 216 "paired" Chinese and English reports on the same events from the 12 days' data. Due to the fact that the frequency distributions of the pageview counts of the paired Chinese/English top news stories were also markedly skewed, the Spearman's Rho test was applied to these paired stories on the same events, and the P value ($P=0.4$; correlation coefficient=-0.135) for this statistical test far

exceeded the 0.05 significance level. We concluded that there was no significant relationship between the pageview numbers of the Chinese and English news reports on the same events. Therefore, we could not estimate the English news reports' ranking or popularity based on their Chinese counterparts' pageview numbers or vice versa. It was safe to say that the Chinese and English readers had very different reading interests while browsing the *People's Daily Online*, which mirrors the findings addressing the first research question (e.g., the Chinese and English readers' different interests in Taiwan and Tibet issues).

Another interesting finding revealed from the "paired" Chinese/English news reports study of this thesis is that the Chinese and English readers' interests in certain political events were dramatically different: the ranking for English news reports on the Chinese leaders' meeting with foreign dignitaries was generally much higher than the ranking for their Chinese counterparts, though these two kinds of news stories were of equal importance in the editor's mind, and placed in the most prominent sections (i.e., the A or B section of the Chinese homepage and the A section of the English homepage), thus having similar visibility on the homepages.

There might be some explanations for this interesting phenomenon. First, the Chinese audience had viewed its leaders' activities via other media, such as television news shows from the previous day. As such stories were almost the same in all kinds of Chinese media, and would be broadcasted first in the late afternoon or evening's prime time, it is likely that first exposure to a news story would occur through the medium of television or radio. In addition, Chinese people are rushing

home or enjoying dinner at this time of day, and thus it is not convenient for them to browse the Web for news. The *People's Daily Online's* 24-hour-traffic volume graph shows that the majority of the readers tended to visit this news site during the 9am-11am and 1pm-3pm periods, which are the business/working hours for the average Chinese people. Newsreaders visiting the *People's Daily Online* generally would not access the stories they were familiar with.

Second, the Chinese readers might not care about the issues of "less important world players". According to the author's observation, except for the leaders or top officials from the world's big powers, dignitaries of less influential nations might not always attract similar attention from the Chinese online audience, a topic that merits further formal study. So far, it is safe for us to conclude that the English readers of the *People's Daily Online* had different interests compared with the news browsers of the Chinese edition.

5.1.5 – Exploring IP Address Distribution

The relative geographic distribution of the readers' IP addresses is worth noting along with the language vs. interest issues. It is not surprising that for the *People's Daily Online* (Chinese Edition), the majority of the IP addresses (85%) were limited to within China during the data collection period (see Figure 5-10 on p.104 for details). However, the majority of English readers' IP addresses (67%) were also from China (mainland) during the same period. This interesting distribution pattern might be due to the increasing number of foreign students/visitors/businessmen or English learners in China. In addition to that, the

second ranking country of the Chinese and English IP address distributions was the same: the United States (see Figure 5-11 on the next page for details). There might be some explanations for this phenomenon: First, more and more people from Mainland China live, work or study in the United States, and they want to obtain the most trusted information on China from the Web. Second, many English-speaking U.S. Web users (e.g., the researchers) also want to read something about China from the most trustworthy official channel.

This study of IP address distribution as well as the English readers' favor of world political issues found by this thesis, might suggest that the *People's Daily Online* should not be named the so-called "Window of China". We might even say that the *People's Daily Online* might be the readers' alternative window to nations other than China. Once the majority of the English edition's visits are generated from the United States and other countries around the world, the *People's Daily Online* might deserve the name "Window of China", which will strengthen the "Voice of China".

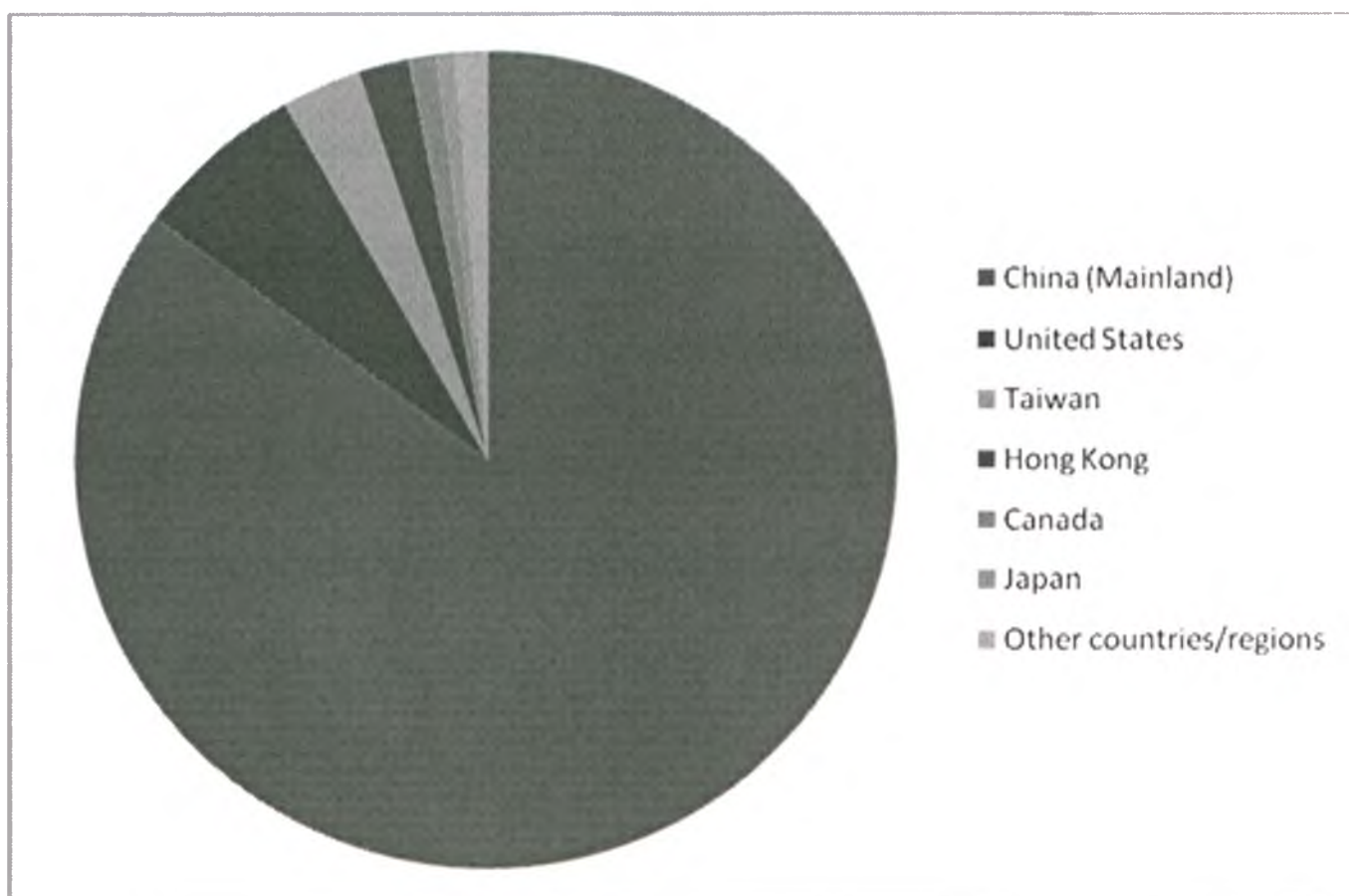


Figure 5-10: Readers' IP Address Distribution for the Chinese Edition

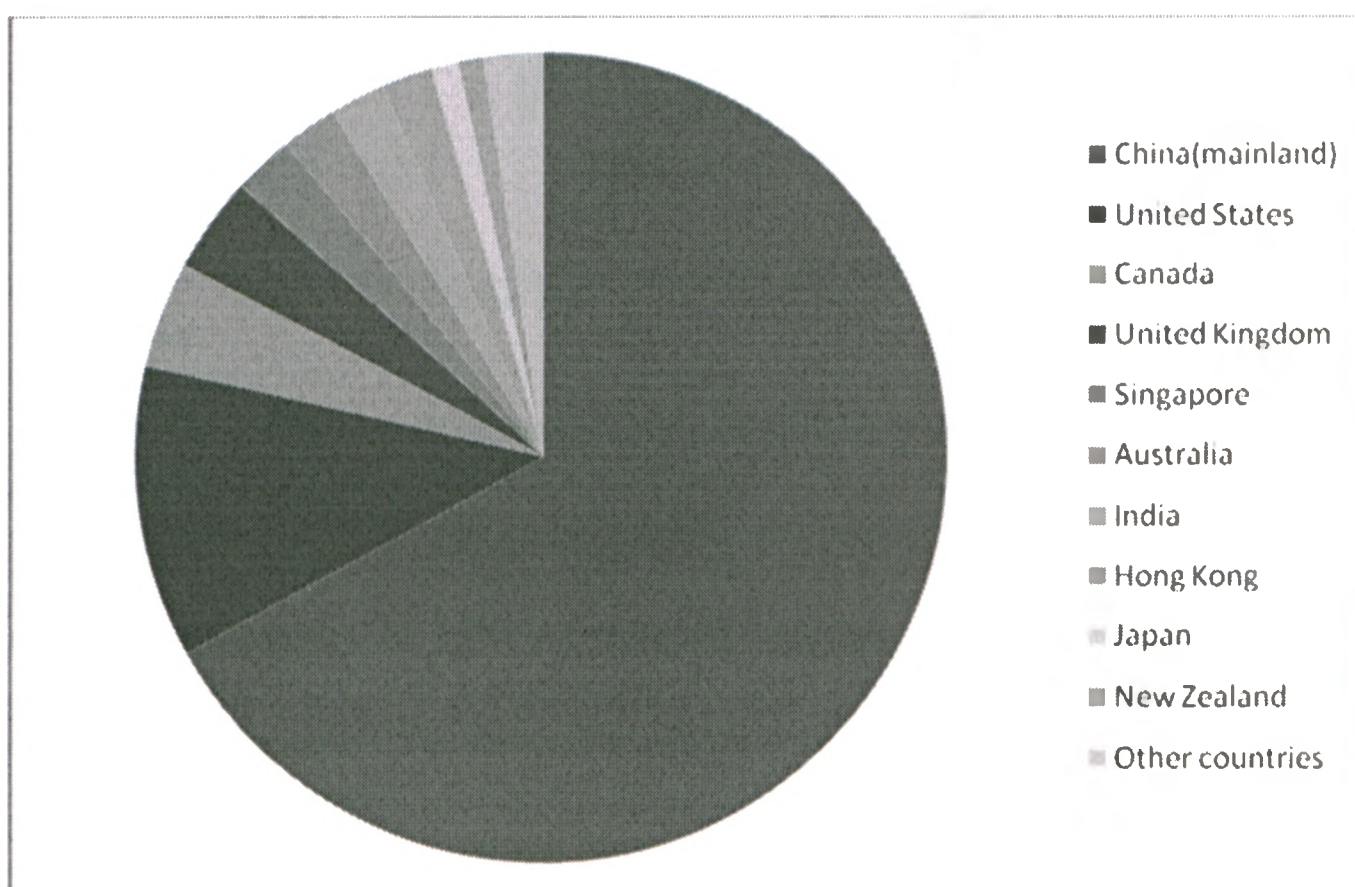


Figure 5-11: Readers' IP Address Distribution for the English Edition

5.1.6 – Traffic Graph to the *People's Daily Online* by Alexa

There is another “spotlight” for us to discuss: the online traffic trend graph drawn by Alexa.com, which is matched with the pageview fluctuation patterns of the *People's Daily Online* (the whole site) over the one-year data collection period (from October 2007 to September 2008); note that peaks in the Alexa traffic graph can always be explained by breaking news or important events (the following graph was generated by Alexa for the traffic of People.com.cn from October 2007 to September 2008).

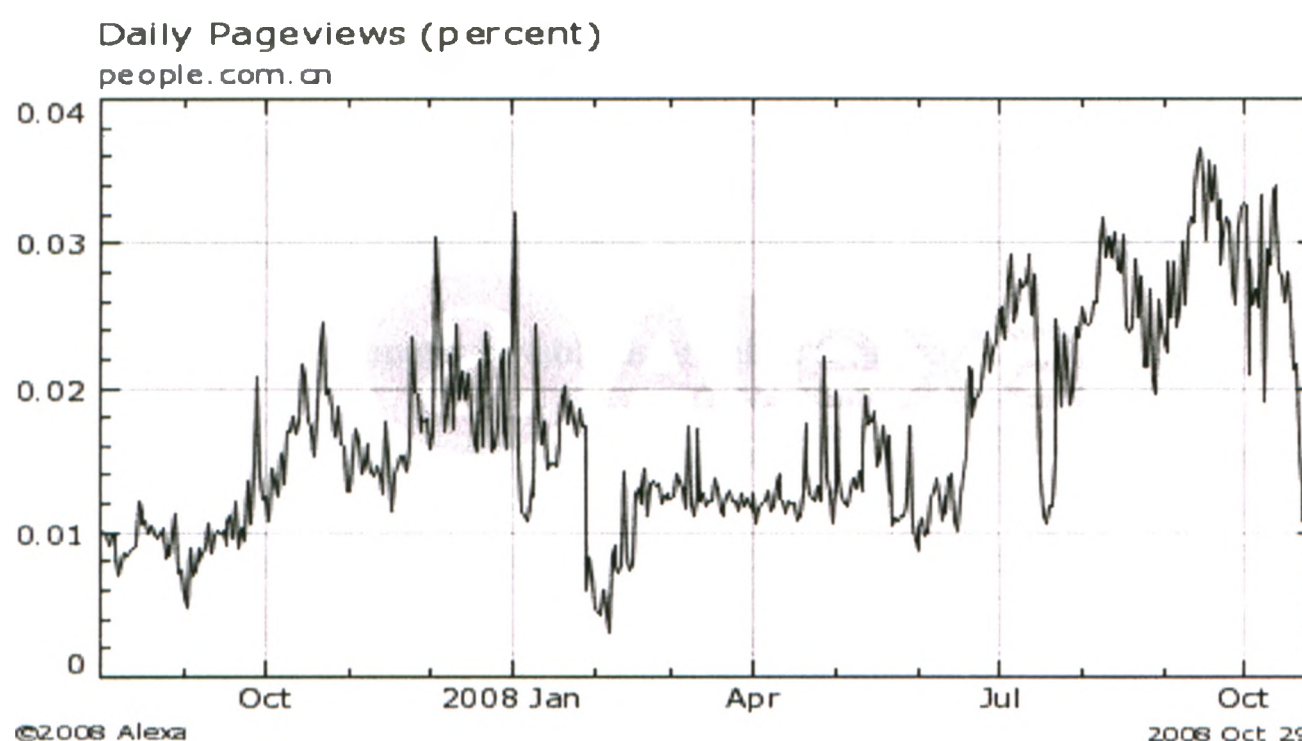


Figure 5-12: Traffic Graph of the *People's Daily Online* Generated by Alexa.com

For October 2007, the graph showed two peaks: one was on the day of the opening ceremony of the 17th CPC National Congress, and the other was on the day of revealing the new CPC leadership, which was in line with the traffic pattern of the *People's Daily Online* during those days. From November to December 2008, many important events happened across China and around the world (e.g., the

Japanese Prime Minister's first official visit to China in several years), which attracted many more visits to the *People's Daily Online*. The traffic receded to one of its lowest points during the Spring Festival (a national holiday for the Chinese) in February 2008, then rebounded during the annual NPC sessions in March, and continuously climbed to another high point immediately after the Sichuan Earthquake. Once there was no big news happening, the traffic volume generally receded as illustrated in the Alexa graph.

In June and July 2008, several incidents of unrest in southwest China's Guizhou Province and the killing of six policemen in their offices in Shanghai stimulated the increasing of online traffic to the *People's Daily Online*, since readers wanted to learn more facts about these issues from the government rather than from rumors or fabricated reports by tabloids or online forums.

The Olympic Games in August 2008 pushed online traffic to a record high: the opening ceremony and Liu Xiang's quit at the first round of the hurdler race generated two traffic peaks. Due to some terrorist attacks that happened during the Olympic Games and the passing away of former Chinese Leader Hua Guofeng, August became "peak show", and even reversed the normally receding trends of traffic volume during the ongoing Olympic Games. In September 2008, visit volumes to the *People's Daily Online* were even heavier than the traffic during the Olympic Games. The tainted baby formula scandal and the launch of "Shenzhou-7" pushed the total visits to the highest yearly point. Since the "Shenzhou-7" was launched one month ahead of its previously scheduled time, there might have been

some considerations from the Chinese authorities to "dilute" side effects or deflect the public attention from the baby formula scandal to "Space Explorations".

The close match of Alexa's traffic graph and the *People's Daily Online's* server log data suggests that the Alexa graph, a publicly available traffic data source, reflects the *People's Daily Online's* traffic pattern very well and thus could be an alternative data source when server log data are not available.

Endnotes

Alexa.com is an Internet company founded in 1996. Alexa's traffic rankings are based on the usage patterns of Alexa Toolbar users and data collected from other sources over a rolling 3 month period. A site's ranking is based on a combined measure of reach and pageviews. Reach is determined by the number of unique Alexa users who visit a site on a given day. Pageviews are the total number of Alexa user URL requests for a site, which were used to draw a trend graph for this thesis. However, multiple requests for the same URL on the same day by the same user were counted as a single pageview.

Alexa's traffic rankings are for top level domains only (e.g., people.com.cn). This company does not provide separate rankings for pages within a domain (e.g., <http://www.people.com.cn/page.html>) or subdomains (e.g., English.people.com.cn) unless they are able to automatically identify them as personal homepages or blogs, like those hosted on Geocities and Tripod (Alexa, 2009).

5.2 – Online Survey Results Analysis

5.2.1 – The 2008 NPC Online Survey Results

A total of 45,070 readers cast 390,300 ballots to this online survey of 30 choices, which were listed in Table 5-17 in descending order of their vote counts.

All options in Survey 2008 were classified in accordance with the same classification scheme for the news reports.

Table 5-17: Distribution of 2008 Survey Votes

Survey Option	Category/Topic	Votes
Judicial Justices	Chinese Politics	24696
Employment Issues	Business	24546
Anti-Corruption	Chinese Politics	23856
Price Rising	Business	23758
Health Care Reform	Health	23078
Income Distribution	Business	22393
Social Insurance	Chinese Politics	21303
Protecting Workers' Rights	Chinese Politics	20971
Education Equality Issues	Education	19090
Housing Issues	Business	16193
People's Political Rights	Chinese Politics	14188
Central Governmental Reform	Chinese Politics	12535
Improving Traffic	Business	12310
Food and Drug Safety	Health	11787
Transparency of Government Information	Chinese Politics	9811
Coordinating Development in Urban and Rural Areas	Business	8386
Population Issues	Chinese Politics	8048
Government Performance Evaluations	Chinese Politics	7894
Eco-Civilization	Environment	7868
Scientific Innovations	Sci-Tech-Edu	7815
Land Protection	Environment	7657
Cell-phone Service Fee	Business	7623
Beijing Olympics	Olympics	7510
Reducing Emissions	Environment	7508
Emergency Reaction	Accident/Disaster	7480
Financial Security	Business	7148
China's "Soft Power"	Chinese Politics	7046
Taiwan Elections	Taiwan	6405
New Officials	Chinese Politics	6088
Industrial Production Safety	Business	5309
Total		390300

Options of Survey 2008 belonging to the same category were further merged into eight categories and listed as “Survey Option Category” in Table 5-18. From the server log statistical data from March 2008 (see Table 5-19 for details), the “target time” for this survey, we found the top reports’ distribution frequencies for the eight matching categories, and also put them into Table 5-18.

Table 5-18: Merging the 2008 Survey Options belonging to the Same Category

Survey Option Category	Votes	Distribution of Top 50 News Chinese Stories in March 2008
Chinese Politics	156436	527
Business	127666	232
Health	34865	69
Sci-Tech-Edu	26905	56
Environment	23033	21
Olympics	7510	15
Accident/Disaster	7480	20
Taiwan	6405	62
Total	390300	1002

Table 5-19: Distribution of Top 50 Chinese Reports in March 2008

News Report Category	Frequency	Percentage
Chinese Politics	527	34.0%
Business	232	15.0%
Entertainment	139	8.9%
Tibet	129	8.3%
World Politics	116	7.5%
Culture/Life/Society	97	6.3%
Health	69	4.5%
Taiwan	62	4.0%
Sci-Tech-Edu	56	3.5%
Sports	33	2.1%
Olympics	20	1.3%
Accident/Disaster	21	1.4%
Crime	18	1.2%
Environment	15	0.9%
Odd	12	0.8%
Other	4	0.3%
Total	1550	100%

5.2.2 – The 2009 NPC Online Survey Results

A total of 150,036 readers cast 946,632 ballots to this online survey of 20 choices, which were listed in Table 5-20 in descending order of their vote counts. All options in Survey 2009 were also classified in accordance with the same classification scheme adopted for the top 50 news reports. Options of Survey 2009 belonging to the same category were further merged into seven categories and listed as “Survey Option Category” in Table 5-21.

Table 5-20: Distribution of 2009 Survey Votes

Survey Option	Category/Topics	Votes
Environment Protection	Environment	83662
Anti-Corruption	Chinese Politics	82517
Food and Drug Safety	Health	77116
Health Care Reform	Health	75478
Income Distribution	Business	70434
Employment Issues	Business	67579
Housing Issues	Business	46715
Education Issues	Sci-Tech-Edu	45661
Social Insurance	Business	44560
Judicial Justices	Chinese Politics	43368
Rule of Law	Chinese Politics	37349
Social Stability	Chinese Politics	35518
Agriculture Issues	Business	34919
Local Government Reforms	Chinese Politics	33348
China's Stock Market	Business	31974
Financial Crisis	Business	31403
Industrial Production Safety	Business	28213
Coordinating development in urban and rural areas	Business	28103
Disaster Relief	Accident/Disaster	24904
Cultural Innovation	Culture/Life/Society	23811
Total		946632

From the server log statistics from March 2009 (requested by the author in 2009), the “target time” for this survey, we found the top news item distribution frequencies for the seven matching categories, and put them into Table 5-21.

Table 5-21: Merging the 2009 Survey Options belonging to the Same Category

Survey Option Category	Votes	Distribution of Top 50 News Reports in March 2009
Business	383900	347
Chinese Politics	232100	333
Health	152594	109
Environment	83662	73
Sci-Tech-Edu	45661	121
Accident/Disaster	24904	35
Culture/Life/Society	23811	46
Total	946632	1064

5.2.3 – Server Log Statistics vs. Survey Results

Since the top 50 news reports' distribution and survey results distribution from March 2008 were highly skewed, a Spearman's Rho test was applied to calculate the correlation coefficient between the two sets of data. Results showed that there was significant correlation between the "top news reports' distribution frequencies" of the eight matching categories and "survey votes" ($P=0.037$; Correlation Coefficient=0.738) in the year 2008.

Another round of the Spearman's Rho test was employed to examine the correlation coefficient between the top news reports' distribution and the survey responses distribution from March 2009, and it also revealed significant correlation ($P=0.014$; Correlation Coefficient=0.857). Therefore, we can conclude that the pageview measure did reflect readers' interests as shown through the survey.

Considering the top queries data of March 2008 that are also available (see Table 5-22 for details), the author adopted a similar method to test the correlation coefficient between the survey votes and query frequency distribution. The

Spearman's Rho test found that there was also a significant relationship between the survey votes and query frequency distribution (Correlation Coefficient=0.762, $P=0.028$). Such results further support the findings presented previously in this chapter, underlining the previous conclusion of this thesis that "server log statistical measures did reflect readers' interests as shown through the survey."

Table 5-22: Distribution of Top 50 Chinese Queries in March 2008

Query Category	Frequency	Percentage
Chinese Politics	467	30.10%
Business	301	19.40%
Entertainment	180	11.60%
Tibet	105	6.80%
World	84	5.40%
Sci-Tech-Edu	81	5.20%
Health	77	5.00%
Culture/Life/Society	65	4.20%
Taiwan	53	3.50%
Sports	51	3.30%
Crime	20	1.20%
Environment	19	1.20%
Olympics	17	1.10%
Accident/Disaster	15	1.00%
Other	15	1.00%
Total	1550	100%

5.2.4 – Survey 2008 and Survey 2009

With two consecutive years' online survey results in hand, we compared the survey votes of 2008 and 2009, so as to see the changing of readers' interests over the one year period. From Table 5-23, we noticed that the total votes and each survey option's votes in 2009 were almost three times higher than those of the 2008

survey. Note that due to the change of political atmosphere and the ending of the Beijing Olympic Games, Taiwan and Olympics issues were not included in the 2009 survey.

Table 5-23: 2008 and 2009 Survey Results

Survey Option Category	2008		2009	
	Votes	Percentage	Votes	Percentage
Business	127666	32.7%	383900	40.6%
Chinese Politics	156436	40.1%	232100	24.5%
Health	34865	8.9%	152594	16.1%
Environment	23033	5.9%	83662	8.8%
Sci-Tech-Edu	26905	6.9%	45661	4.8%
Accident/Disaster	7480	1.9%	24904	2.6%
Culture/Life/Society	N/A	N/A	23811	2.5%
Olympics	7510	1.9%	N/A	N/A
Taiwan	6405	1.6%	N/A	N/A
Total	390300	100%	946632	100%

The business issues received more votes than the Chinese political ones (40.6% vs. 24.5%) in the 2009 survey, which might be reflective of the changing of readers' interests over a year with a bleak global economic outlook. In 2008, the percentage of votes for the business and Chinese political issues were 32.7% and 40.1% respectively, illustrating that the public cared more about domestic political events in 2008. The tainted baby formula scandal that happened in September 2008 likely accounted for the considerable rise in the Chinese people's attention to health issues in 2009 (16.1% of the total votes) as compared to 2008 (8.9% of the total votes). The Chinese government had been advocating "the Green Olympic Games" from the beginning of 2008, which made the public realize the importance of environment protection, and the percentage of votes for such issues increased from

5.9% to 8.8%. The Sci-Tech-Edu topics attracted relatively fewer votes in the 2009 Survey (4.8%) than in 2008 (6.9%) because there were few widely expected scientific breakthroughs this year. For example, no achievement like the first Chinese space walk in 2008 is likely to emerge in 2009, and therefore might explain why the public voted less for Sci-Tech-Edu issues in this year.

5.3 – Analyzing Inlinks to the *People's Daily Online*

A total of 200 Chinese and 200 English pages with inlinks to the *People's Daily Online* were analyzed. These inlinking pages were randomly sampled from every five results of the top 1000 URLs retrieved by Yahoo!. Compared with the classification scheme for the top news items, not all of the news report categories were inlinked by the retrieved websites (e.g., entertainment news reports). Besides frequently updated news items, some pictures and “static” information offered by the *People's Daily Online*, such as the organization of the Chinese government, the Full Text of China's Constitution, and a general introduction to China's provinces, were among the top inlinked pages.

5.3.1 – Inlinks Profile of the *People's Daily Online*

For the webpages with inlinks to the Chinese content, all of the retrieved webpages were in Chinese, with about 86.5% (173) from China, and 9% (18) from the United States, while for the pages with inlinks to the English content, about 73.5% (147) were from the United States, and 5% (10) from the U.K. (see Table 5-24 for details).

Table 5-24: Originating Countries of the Inlinks

Country	Inlinks to Chinese Content	Inlinks to English Content
China	173 (86.5%)	7 (3.5%)
U.S.	18 (9%)	147 (73.5 %)
Canada	2 (1%)	6 (3%)
U.K.	2 (1%)	10 (5 %)
Other	5 (2.5%)	20 (15%)
Total	200 (100%)	200 (100%)

The inlink counts were fairly stable from May to October 2008, during which period the author collected the data every four weeks. After comparing the first and last 100 retrieved Chinese inlinking pages, the author noticed that the distribution of the categories of the inlinked news reports/information, as well as the purposes of the inlinking between the two groups, were almost the same. Similar patterns also existed in the English inlinking pages retrieved by Yahoo!.

There was one interesting thing for the inlinking pages: more than 5% (10) of the sampled inlinks to the English URLs within the domain of the *People's Daily Online* claimed they were linking to the webpages of the "Xinhua" or "China Daily", two other leading English news providers in China. Meanwhile, such errors were rare for the Chinese inlinking pages, and the author only found one out of the 200 sampled inlinks.

Table 5-25 shows the content (i.e., news report categories) breakdown of inlinked pages of the *People's Daily Online*. When English and Chinese are merged (see last column of this table), Sci-Tech-Edu reports (18.75%), World Politics news stories (15.75%), Business news reports (13%) and Chinese Politics stories (12.25%) were the popular contents being inlinked.

Table 5-25: Contents of the Inlinked Pages

Inlinked Category	Inlinked Pages' Language		Total (Percentage)
	Chinese (Percentage)	English (Percentage)	
Sci-Tech-Edu	39(19.5%)	36(18.0%)	75(18.75%)
Chinese Politics	38(19.0%)	11(5.5%)	49(12.25%)
Business	26(13.0%)	26(13.0%)	52(13.00%)
BBS	24(12.0%)	0(0%)	24(6.00%)
Pic published by PD	24(12.0%)	3(1.5%)	27(6.75%)
Culture/Life/Society	22(11.0%)	13(6.5%)	35(8.75%)
Sports	10(5.0%)	9(4.5%)	19(4.75%)
Environment	5(2.5%)	12(6.0%)	17(4.25%)
Health	5(2.5%)	11(5.5%)	16(4.0%)
Data offered by PD	3(1.5%)	11(5.5%)	14(3.50%)
Other	3(1.5%)	0(0%)	3(0.75%)
World Politics	1(0.5%)	62(31.0%)	63(15.75%)
Accident/Disaster	0(0%)	6(3.0%)	6(1.50%)
Total	200	200	400

A statistically significant relationship was found between the languages of inlinked pages and the contents of inlinked pages (Chi-square test, $p < 0.01$). The more commonly inlinked Chinese news reports were Chinese Politics (including the Taiwan and Tibet issues) stories (19.0%), contents from the BBS (12%), and pictures from the *People's Daily Online* (12%), while the English sites linked more World Politics reports (31%), environmental reports (6%), health care issues (5.5%), static information (5.5%), and accident/disaster stories (3.0%). This is in line with our previous analysis of the server log statistics which also provided evidence that Chinese readers accessed more Chinese political stories and that English readers browsed more World political reports. Other categories of news, such as Business and Sci-Tech-Edu, got almost the same amount of coverage from the Chinese and

English inlinking webpages. These results showed us the differences in Chinese/English readers' uses of information from the *People's Daily Online* as reflected in the inlinks' distribution patterns.

Reasons or purposes for linking to the Chinese and English editions of the *People's Daily Online* are summarized in Table 5-26. When the two languages were combined, we found that "Indicating source of news" accounted for the majority of the inlinks to the *People's Daily Online* (57.5%), which told us this online newspaper was used mainly as news source. A statistically significant relationship was found between the language (Chinese or English) of the inlinked contents and the reasons or purposes for linking to them (Chi-Square Test, $p < 0.01$). For the inlinks to the Chinese contents from the *People's Daily Online*, the purpose of inlinking was more likely to be "Providing further information", while the purpose for the English inlinks was more likely to be "Indicating source of news". Such phenomena told us that the *People's Daily Online* was mainly used by English readers as an information source.

Table 5-26: Purposes of Creating Inlinks

	Inlinking Purpose		Total
	Indicating source of news	Providing further information	
Inlinks to Chinese Pages	72 (39%)	128 (61%)	200
Inlinks to English Pages	158 (79%)	42 (21%)	200
Total	230 (57.5%)	170 (42.5%)	400(100%)

"Indicating source of news" here referred to the inlinked URLs that had direct ties with contents of the inlinking pages. For example, the news sites used inlinks to

direct readers to the sites containing full text of the story. The webloggers also used inlinks to tell the readers where his/her words came from. "Providing further information" denoted URLs that did not have direct ties with the inlinking pages' contents; in other words, they were used by the creators of the inlinking pages as supplementary information resources. Such inlinks could be added or removed without affecting the integrity of the inlinking pages or raising copyright concerns.

Webpages linking to the *People's Daily Online* were classified into various types as discussed earlier in the Data Collection Methodology section (Chapter 4.3.3) of this thesis. The classification results are summarized in Table 5-27. The most common type of sites was "weblogs" (28.5%), followed by "corporate sites" (19%) and "news sites" (17%). It is somewhat surprising that "organizational sites" (9.75%) and "personal sites" (9%) were more common than other sites.

A statistically significant relationship was found between the language of inlinked pages and the site type of linking pages (Chi-square test, $p < 0.01$). As shown by the percentage figures in Table 5-27, the Chinese Edition attracted more inlinks from corporate sites, BBS sites and online portals (31.5%, 6.5% and 3.5% respectively), while the English Edition attracted much less inlinks from these site types (6.5%, 0% and 0% respectively). On the other hand, the English Edition attracted relatively more inlinks from Weblogs (31%), news websites (24%), and personal sites (15.5%). Such phenomenon showed the differences in Web applications: Chinese readers used more BBS sites, while English readers more opted for personal sites. The Weblogs got similar usage from Chinese and English

readers. The Chinese Edition attracted more inlinks from corporate and governmental sites, while the English Edition attracted more inlinks from organizational sites. All these phenomena reflect the Chinese Edition's "most trustworthy news resource" status in China (CASS, 2002), as well as the English Edition's popularity among personal Web users in the English speaking community.

Table 5-27: Site Type of Inlinking Pages

Site Type	Chinese	English	Total (Percentage)
BBS	13(6.5%)	0	13 (3.25%)
Com	63(31.5%)	13(6.5%)	76 (19%)
Edu	9(4.5%)	7(3.5%)	16(4%)
Gov	14(7.0%)	3(1.5%)	17(4.25%)
News Media Website	8(4.0%)	6(3.0%)	14(3.5%)
News Website	20(10.0%)	48(24%)	68(17%)
Online Portal	7(3.5%)	0	7(1.75%)
Org	9(4.5%)	30(15%)	39(9.75%)
Personal	5(2.5%)	31(15.5%)	36(9%)
Weblog	52(26.0%)	62(31.0%)	114(28.5%)
Total	200	200	400

5.3.2 – Inlinks and Readers' Interests

Table 5-25 lists the content of the inlinked pages from the *People's Daily Online*. Was there any relationship between the "most inlinked news reports" and the "most visited news reports" listed in Table 5-2 and Table 5-3? Since the two data sets were highly skewed, the Spearman's rho test was employed to examine the relationship between them. For the English news stories, there was a significant relationship between the most visited news reports and the most inlinked news

stories ($P=0.026$, Correlation Coefficient=0.693). For the Chinese news stories, there was no significant relationship between the most visited news reports and the most inlinked news stories ($P=0.16$, Correlation Coefficient=0.480).

It must be noted that only the ten matching categories were compared, including Sci-Tech-Edu, Chinese Politics, Business, Culture/Life/Society, Sports, Environment, Health, World Politics, Other, and Accident/Disaster news reports. Because we could only sample the top 1000 inlinks retrieved by Yahoo!, results from this round of Spearman's rho test might not be complete. At least for the collected data, the inlinked English pages within the domain of the *People's Daily Online* could have reflected readers' interests measured by the pageview numbers, which was in line with the previous finding that English readers used more contents from the People's Daily Online than their Chinese counterparts (see Table 5-26 on p.112 for details).

To further study inlinks to the *People's Daily Online*, we retrieved inlinks to each Chinese news channel using Chinese Yahoo's "linkdomain: channeldomain – site:sitedomain" query (see Table 5-28 for the sample query). Since the linkdomain command could not retrieve total inlink counts to the English news channel at the time of data collection (September 2008), we had to use the "link:pageURL – site:sitedomain" query to collect the inlinks to the homepage of each English news channel as an alternate (see Table 5-28 for the sample query). The retrieved inlink counts to the individual channels were listed in Table 5-29.

Table 5-28: Queries to Collect Inlinks to the Chinese/English News Channel

Channel Name	Chinese	English
Chinese Politics	linkdomain: politics.people.com.cn -site:people.com.cn	link:http://english.people.com.cn/ 90001/90776/ -site:people.com.cn
World Politics	linkdomain: world.people.com.cn -site:people.com.cn	link:http://english.people.com.cn/ 90001/90777/ -site:people.com.cn

Table 5-29: Inlink Counts to the Chinese/English News Channels

Chinese Channel	Inlink Count	English Channel	Inlink Count
Education	286,000	Chinese Politics	12,400
Health	184,000	World Politics	8,290
Taiwan	183,000	Sports	2960
Business	168,000	Business	997
Entertainment	158,000	Sci-Edu	647
Sports	120,000	Life	262
Chinese Politics	108,000		
Life	101,800		
Sci-Tech	71,400		
World Politics	48,900		
Military	43,300		

Both the inlinks and the server log statistics were highly skewed; thus the non-parametric Spearman's Rho test was employed here. With the current data in hand, one disappointing thing about the inlink analysis for this thesis was we could not establish significant ties (Spearman's Rho test, $P > 0.1$, see Table 5-30 for details) between the inlink counts to the Chinese/English news channels and the latter's total pageview numbers at the time of data collection; thus, the external inlink counts might not be a good alternative indicator of readers' interests.

Table 5-30: Relations between Inlink Counts and Channel Pageview Numbers

	Correlation Coefficient	P-Value
Chinese	0.006	0.987
English	0.600	0.208

5.3.3 – Interlinking Structure among Leading News Websites

According to the author's personal observations since 2002, the homepages of popular English online newspapers contain few outgoing hyperlinks to other online newspapers or media agencies, but their websites do heavily interlink with each other via links from reports and weblogs (see Table 5-31 for details).

Table 5-31: Interlinks among leading U.S. online newspapers

	New York Times	Washington Post	USA Today	Wall Street Journal
New York Times to		15,300	43,100	80,700
Washington Post to	41,800		17,600	14,300
USA Today to	35,600	17,800		9,690
Wall Street Journal to	31,300	14,400	3,260	

Contrary to this, homepages of the Chinese online news agencies always have outlinks to their competitors. For example, the *CCTV* and the *People's Daily Online* are connected with each other by tens of thousands of links (see Table 5-32 for details).

Table 5-32: Interlinks between the *People's Daily Online* and its Chinese Partners

	China Daily	CCTV	CNRadio	Xinhua
From People.com.cn to	60,900	68,400	92,400	439,000
To People.com.cn from	4670	34,400	87,600	57,200

All link counts in Table 5-31 and Table 5-32, as well as Table 5-33 and Table 5-34 in the coming sections, were collected in January 2009, with the help of Yahoo!'s linkdomain command. For example, we could retrieve the inlinks to the *People's Daily Online* from the *New York Times*, with the help of the following sample query: **linkdomain:people.com.cn site:nytimes.com.**

The interlinking data between the *People's Daily Online* and other major official news websites in China showed that the former contributed more inlinks to latter, than what it received from them (the only exception is the *CNRadio*, China's national broadcast service). Considering the Chinese online news agencies generally created outlinks to the homepage or the pages where the reprinted stories come from, such interlinking data might reveal the flow of exclusive reports among various Chinese websites. In our cases, news stories from the *Xinhua News Agency*, *China Daily*, and *China Central Television* (CCTV) were more likely to be reprinted by the *People's Daily Online* than were stories from the *People's Daily Online* reprinted by the other three news websites.

The interlinking structures between the *People's Daily Online* and the four leading U.S. newspapers (the *New York Times*, *USA Today*, the *Wall Street Journal* and the *Washington Post*) listed in Table 5-33 also need to be mentioned. From Table 5-33, we can easily find that the *People's Daily Online* received overwhelmingly more inlinks from the four leading U.S. newspapers' websites than what these overseas online newspapers received from their Chinese counterpart, which were all created for the purpose of "Providing further information".

Table 5-33: Interlink Counts between *People's Daily Online* and U.S. Newspapers

	New York Times	Washington Post	USA Today	Wall Street Journal
To People.com.cn from	1,210	108	248	80
From People.com.cn to	8	11	7	15

To further study the interlinking phenomenon between the *People's Daily Online* and the four U.S. online newspapers, the author did a content analysis of the inlinks from these newspapers to the *People's Daily Online*. Forty inlinks to the *People's Daily Online* from each of the four U.S. newspapers were randomly chosen and classified in accordance with the previously mentioned classification schemes (news story categories and linking purposes) in Chapter 4.3.3.

A total of 160 inlinks were analyzed. Table 5-34 shows that the Chinese Politics (30.6%), the World Politics (26.9%) and the Business (16.9%) were the top three categories of news reports linked by the U.S. media. When the four newspapers were examined separately, we noticed that the *New York Times* linked more Chinese Politics contents (40%) from the *People's Daily Online* than the *Wall Street Journal* (37.5%), the *Washington Post* (37.5%) and *USA Today* (7.5%). The *Washington Post* linked more World Politics news stories (55%) from the *People's Daily Online* than *USA Today* (45.0%), the *New York Times* (5%) and the *Wall Street Journal* (2.5%). The *Wall Street Journal* linked more Business (37.5%) news stories than *USA Today* (22.5%) and the *New York Times* (7.5%).

Table 5-34: Interlinking between *People's Daily Online* and U.S. Newspapers (1)

News Category	Newspaper				Total
	New York Times	USA Today	Wall Street Journal	Washington Post	
Chinese Politics	16(40%)	3(7.5%)	15(37.5%)	15(37.5%)	49(30.6%)
Culture/Life/Society	5(12.5%)	0	1(2.5%)	0	6(3.8%)
Business	3(7.5%)	9(22.5%)	15(37.5%)	0	27(16.9%)
Health	2(5%)	0	0	0	2(1.3%)
Olympics	7(17.5%)	7(17.5%)	3(7.5%)	0	17(10.6%)
Sci-Tech-Edu	5(12.5%)	3(7.5%)	5(12.5%)	3(7.5%)	16(10.0%)
World	2(5.0%)	18(45.0%)	1(2.5%)	22(55%)	43(26.9%)
Total	40	40	40	40	160

The interlinking reasons/purposes between the U.S. newspapers and the *People's Daily Online* are summarized in Table 5-35, which tells us that "Indicating source of reports" (81.3%) was the predominant linking purpose for the four U.S. newspapers as a whole. Once we looked at each newspaper, 40% of the inlinks generated by the *New York Times* for "Indicating source of reports" and 60% of inlinks for "Providing further readings". The purpose of creating inlinks from the other three U.S. newspapers to the *People's Daily Online* was overwhelmingly "Indicating source of reports".

Table 5-35: Interlinking between *People's Daily Online* and U.S. Newspapers (2)

Newspaper	Linking Reason/Purpose		Total
	Indicating Source of Reports	Providing Further Readings	
New York Times	16(40.0%)	24(60%)	40
USA Today	40(100%)	0	40
Wall Street Journal	36(90%)	4(10%)	40
Washington Post	38(95%)	2(5%)	40
Total	130(81.3%)	30(18.8%)	160

Something also worth noting is, except for outlinks indicating source of report following the title of each news-page, the *People's Daily Online* and other leading Chinese news websites never embedded outlinks to their so-called “partners” in the news reports published within their domains. Such phenomena showed that the Chinese online news media were actually reluctant to link with each other.

5.4 – Examining the Top 50 News Reports’ Online Reprints

The top 50 Chinese/English news reports’ pageview numbers are highly skewed (see Figures 5.1 and 5.2 for the histograms); therefore, the Spearman's rho test was employed to investigate the relationships between the top news reports’ pageview numbers and their online reprint rates returned by English and Chinese Google from the 12 randomly selected days. There was no significant relationship between the pageview numbers and online reprint rates for each of the top 50 English reports. However, the online reprint rates for each of the top 50 Chinese news items were positively correlated with the latter’s pageview numbers (see Table 5-36 for details).

Table 5-36: Ties between Pageview Numbers and Online Reprint Rates

Report Type	Correlation Coefficient	P-Value	Sample Size
English Reports	0.012	0.776	600
Chinese Reports	0.136	0.001	600

It seems that the online reprint rates for the Chinese top 50 news items might be a publicly accessible alternative data source for us to investigate the readers' interests when server log data are not available. However, the Correlation Coefficient for the Chinese test was too small (0.136), and did not support a strong relationship.

To further study this issue, the author conducted several rounds of correlation coefficient tests at the group (news report's category) level and individual report level. Both of the sampled 600 English news items and 600 Chinese reports were divided into 14 groups in accordance with the classification scheme discussed earlier in Chapter 4. The total pageview number of each group was calculated by SPSS.

Because the pageview numbers were markedly skewed at the group level, the Spearman's Rho test was employed to compute the correlation coefficient between the pageview numbers and online reprint rates at the group level. Both the English and Chinese reports' pageview numbers were significantly correlated with their pageview numbers at the group level (see Table 5-37 for details).

Table 5-37: Pageviews vs. Online Reprints at the Group Level

Report Type	Correlation Coefficient	P-Value	Sample Size
English Reports	0.917	P<0.001	17
Chinese Reports	0.899	P<0.001	17

However, correlation tests carried out within each news group (i.e., individual news items as the unit of data analysis) found no statistically significant relationship between pageview numbers and online reprint rates (the maximum correlation coefficient=0.291, and the minimum P-value=0.08). Considering there was no statistically significant relationship between the pageview number of each of the 600 selected English top stories and their online reprint rates, the author did not conduct the “within each group level” correlation co-efficient test for the English reports.

As publicly accessible data, the online reprint rates of the top 50 Chinese/English reports were significantly correlated with the pageview numbers at the news reports’ group level. So the online reprint rates can reflect the readers’ interests at the group level, and this is true for both the English and the Chinese news reports. However, there was no statistically significantly relationship between the online reprint rates and the pageview numbers of the top news reports at the individual news item level. Thus, ranking the individual news report by online reprints might not reflect the readers’ interests as measured by the news reports’ pageview numbers.

One possible explanation for this issue might be found in the methods of news dissemination: if someone could access the news from other sites, he/she may not

have needed to access the same items again from the *People's Daily Online*; however, this is just a speculation, since we could not retrieve the pageview numbers from the websites that reprinted news from the *People's Daily Online*. Considering some reports do not have an author, and using the keywords from the titles returns too many results, at least for now, searching the online reprints by the "Titles" might not be the best method, but it is the most feasible one, since the full text article is too large to be the query for commercial search engines.

Chapter 6 – Conclusions and Future Research

The Web has changed the media industry and the public's information-seeking behaviors dramatically. The financially troubled and always delayed mail-delivered newspaper business cannot meet demands of the younger generations, who also form the majority of the fast-growing Web users. There is a well-known principle within the journalism circle that newspapers could not survive over a long run if they had to rely on circulation revenues or government subsidies. Although the *People's Daily* will not stop its print publication with the Communist Party as the ruling party of China, it has already begun to take some precautions to increase profits from the non-printing business. Thus, seeking "new economic growth points" from the Web becomes a priority, and enhancing website content is crucial for the editorial board to keep this more than 60-year-old national newspaper "abreast of the times".

Future development of the online newspaper depends on significant growth of Web traffic to its vast free content, which may attract more advertising customers, as well as more subscriptions to its recently introduced paid information services. Such practices might change the fundamentals of Chinese journalism and the economic models underlying the traditional newspaper industry, and requires a thorough readers' interests study for the *People's Daily Online*. This doctoral thesis explored this issue from different perspectives (i.e., server log statistics, surveys, inlinks and online reprints), and has made some contributions to the knowledge of information science and media studies in addition to suggestions to the *People's Daily Online*.

6.1 – Findings of This Doctoral Thesis

In 2008, the Web experienced rapid development and wide application in China. Disasters such as snow storms during the Lunar New Year holiday in early 2008, the massive earthquake that struck China's Sichuan Province in May as well as the Beijing Olympic Games in August continuously stimulated growth in the number of Web news readers and visits to the *People's Daily Online*. Analyzing the server log statistics of this online newspaper as well as publicly accessible data yielded some statistically significant results. According to the daily pageview numbers of the top 50 Chinese/English news items, we learned which categories of reports attracted relatively more online visits. Since pageviews generated by the top 50 news items constituted an overwhelming majority of total visits each day, conclusions made from such data reflected the real situation. Because the top 50 news reports' pageview numbers were significantly correlated with the comment counts and emailed frequencies of the same items, we were also able to rank the top news items by the other two measures, which also reflected readers' interests.

6.1.1 – Chinese and English Readers had Different Interests

This doctoral thesis not only identified issues of common concern for the Chinese/English readers (such as the news reports on Chinese Politics, Business, Olympics and Sichuan Earthquakes) but also identified events generating different reactions (such as Tibet and Taiwan-related stories). In the meantime, there was no relationship between the pageview numbers of the Chinese and English reports on the same events, further revealing the divergence in predominant points of interest

between the different language readers. In the future, a more comprehensive study on news reports in Japanese, Russian, French and Arabic published by the *People's Daily Online* will be necessary to support the findings from the current study.

The author noticed that the rankings of English reports on Chinese leaders' visits to "non-influential" nations or the Chinese leaders' domestic tours were much higher than their Chinese counterparts. In addition, English news on visits of many foreign leaders' to Beijing generally ranked higher than the same reports in Chinese. These two types of lower-ranking Chinese news items did not attain a position in the top 50 news reports with highest pageview numbers, while their English counterparts ranked within the top 10.

Something also worth noting was that stories of the U.S. President or other senior U.S. officials' visits to China/meetings with Chinese leaders ranked high in both the English and Chinese top 50 news reports' lists. All of the above mentioned "Chinese leaders involved" news reports were treated as the most important ones and placed in the prominent locations of the homepages (i.e., Homepage Pic, A or at least B sections on the Chinese homepages; Homepage Pic or A sections on the English homepages but they received mixed reactions from the readers. Due to the limitations of the collected data, which only include the top 50 news items, the author could not conduct formal statistical tests to examine this issue; however, this phenomenon has been earmarked for future studies.

Another interesting finding is the Chinese readers cared about tabloid news reports on Chinese and overseas stars, such as Paris Hilton or Zhang Ziyi, but the

English readers almost exclusively focused on non-Chinese celebrities, since the English stories on Chinese entertainment stars never showed up in the top 50 lists. However, stories on the NBA star Yao Ming and famous Chinese athlete Liu Xiang always have the potential to attain a position in both the English and Chinese top 50 news reports lists, although their rankings would likely be unequal.

Analyzing distribution patterns of the query topics, which more directly mirrored readers' information needs, offered us some promising ideas/clues on the Web users' interests. Correlations between the pageview numbers of the top 50 Chinese/English news and the related query counts were statistically significant. Thus, we can confidently say that the top news reports' lists ranked by pageviews are a very good source to study the online news readers' interests and needs.

This doctoral thesis shed some light on the general interests of the readers visiting the *People's Daily Online*; however, a more detailed analysis of the most popular news reports will be necessary for the channel editors to improve the popularity of their contents. For example, based on the conclusions of this thesis, we should continue to study what kind of Chinese/World political stories will attract more visits than others. The author plans to conduct such "event-based" classifications working along with staff reporters of the *People's Daily Online* in the coming years, with the aim of "predicting" the possible ranking of news reports on other countries, so as to set up better editorial rules for the International News Channel of this website.

6.1.2 – Web Interaction Reflects Public Attitude on Specific Nations

At the present time, this doctoral thesis found that the Chinese news readers were more interested in reading negative news stories on the United States and Japan, since the negative reports attracted more readership as measured by pageview numbers. A similar phenomenon did not occur for Russia-related issues. This conclusion was supported by a qualitative content analysis of the comments posted to the United States/Japan/ Russia-related news reports. Such findings reflected the public sentiments toward the two economic powers and Russia. In other words, findings from the virtual community reflected the Chinese people's attitude in the real world. This demonstrated that Web server log data could be a good source for us to gauge the public opinion on specific domestic and international issues in future studies.

Results from the thesis could play an important role in helping the *People's Daily Online* publish more reports that better meet its readers' needs. However, we must note that putting too many U.S./Japanese-related negative reports on the homepage is not a very professional idea to attract readers, since such stories might mislead the public opinion and stimulate extreme nationalist sentiments in China.

6.1.3 – Editors' Choices DO NOT Match Readers' Needs

One interesting finding by this thesis is that the "old fashioned propaganda techniques" employed by the Chinese media did not work well. The *People's Daily Online* is the official "mouth piece" of the CPC Central Committee on the Web, but reports of the CPC Leaders placed in the "best" section on the Chinese homepage

did not attract more readers, especially for those news titles full of "empty talks" or just political cliché to praise the leaders or state government's "great achievements".

On the English homepage, some news items placed under the Homepage Pic and A sections used to be static for a long time, which did not attract enough pageviews for them to be ranked in the top 50 news list (these did not appear in the top 100 news list collected in October and November 2007 either). However, such news reports were ranked pretty high (top five) on the first day of their publication on the Web. On Sundays or holidays, the English editors keep two homepage photos, or some of the A sections' titles unchanged, hoping such "most popular news stories of yesterday" will continuously attract relatively more visits. Unfortunately, even for those ranked first in the previous day, if they were left on the homepage unchanged, we could not find their trace in the top 50 lists for the following day. This underlined the long-term journalism principle that updated "fresh" contents are crucial for news reports to attract readers; otherwise, they waste the "good positions" on the homepage. The "destiny of the unchanged news reports" may also imply that readers of the *People's Daily Online* English Edition were relatively stable and frequently visited this news website; thus, they did not need to click on the news items they had browsed before. However, this conclusion is just based on the author's personal observations and speculations, and requires further study in order to make any formal conclusion.

For the Chinese homepage, similar findings were also observed: the Chinese readers tended to retrieve more stories on personal health care issues than the

English Web users. Articles on "improving male's sexual ability or female's beauty" always ranked higher than other health care reports, and sometimes even occupied the first place of the top 50 news reports list. These popular articles would be left on the homepage for several days to attract more readers. Like the English top news stories, such "unchanged" articles on the homepage could not maintain their ranking in the top 50 news list for two days. Another finding is that these health care related stories were always placed in the least visible section of the Chinese homepage, but the pageview numbers they received on the first day overrode their "weights" in the editor's mind.

This doctoral thesis proved that for the *People's Daily Online*, the most prominent sections on the homepage plus the "most important news reports" did not yield more visits than the "attractive" news items placed in the less visible sections of the homepages. However, many older-styled English reports (the most important ones in the editor's mind) still drew much attention from readers, and ranked higher in the top 50 news list than their Chinese counterparts.

6.1.4 – Publicly Accessible Data of the News Websites

There is one spotlight from the publicly accessible data retrieved within the news websites being studied – comments posted by readers. There was a significant correlation between the top 50 news reports' pageview numbers and comments counts, and a qualitative analysis of the comments on United States/Japan/Russia-related news items showed that they were direct reflections of the Chinese readers' attitudes towards specific events or nations. Thus, we could employ online

comments as an important data source to study readers' interests in the future.

The other two types of important publicly accessible data – external inlinks and online reprints – told us some different stories. These inlinks had been identified by many researchers as alternative measurement indicators of business or academic performance (Vaughan, 2005). Unfortunately, this doctoral thesis could not find a significant correlation between the inlink counts and pageview numbers of the individual news channels (with the help of the current data in hand). However, inlinked English pages retrieved by Yahoo! within the domain of the *People's Daily Online* did reflect readers' interests as measured by the pageview numbers. Thus, inlinks might still offer us some supplementary information on readers' interests.

The online reprint rates retrieved by Google also deserve our attention. There was no significant correlation between the pageview numbers of top 50 English news items and their online reprint rates. However, we did find a significant correlation between the pageview numbers of the top 50 Chinese news items and their online reprint rates. However, such a relationship was not strong, as reflected in the small correlation coefficient value.

Further analysis showed that there was no statistically significant relationship between the online reprint rates and the pageview number of the top news reports at the individual news item level. Thus, ranking news reports by their online reprint rates did not reflect the readers' interests as measured by the pageview numbers.

The Web information, in particular visits to news website and publicly available data retrieved by commercial search engines, can be very dynamic, which might be the possible explanation to the findings on inlinks and online reprints studies of this doctoral thesis. In the future, we might find some better methods to collect inlinks or online reprints, which might lead to different conclusions.

6.1.5 – The Interlinking Phenomena among News Websites

In addition to the “cited contents” from the news websites indentified by the inlinks, the interlinking structures among the leading news sites in China and abroad are also worth noting. In China, homepage of the *People's Daily Online* has been a portal where editors will point visitors to other areas on the Web. However, the author could not find links from the homepages of four leading U.S. newspapers going to the websites of other news agencies, with the exception of advertising agencies.

The interlinking structures among the *People's Daily Online* and four leading Chinese news media sites showed that the former contributed more inlinks to the latter, than what it received from them, which might reveal the flow of exclusive reports in China. In the meantime, there were many more inlinks to the *People's Daily Online* from the U.S. media sites than there were inlinks from the opposite direction. An analysis of the inlinks from the *New York Times*, the *Washington Post*, the *Wall Street Journal* and *USA Today* to the *People's Daily Online* showed that many were from reports/blogs indicating the source of information, or from the compilation of related stories. And the inlinks from the *People's Daily Online* to the

four U.S. newspaper websites were all for the purpose of “providing further information sources”. Such results may suggest that the U.S. news media are relatively more “internationally open” than their Chinese counterpart – the *People's Daily Online*. Another difference between the Chinese and U.S. leading news media websites was that the former embedded much fewer links to outside resource, especially the other media sites in the body of news report published within their domains.

6.2 – Recommendations Made by the Thesis

This doctoral thesis addressed a classical question of information science: if and how users' information needs are being served, and it did so through analysis of server log statistics from a leading online newspaper in China, the biggest Web/information market in today's world. It examined and provided evidence for the usefulness of server log analysis in studying information needs. Results from this study show that editors' choices and readers' favorites do not always match each other; thus, the content of a news item (i.e. its title on the homepage) is more important than its homepage position in terms of attracting more online visits, a finding that differs from those of Bar-Ilan et al. (2009).

Another finding was that Chinese and English readers' interests in the same events were also different; for example, many English reports of Chinese Top Leaders' activities received much more Web attention than their Chinese counterparts. Since English and Chinese readers have different information needs, any bi-lingual or multi-lingual news website must tailor its content and homepage

structures accordingly.

In early 2009, the author participated several online discussion sessions on the homepage redesigning issues via Windows messenger with some of the staff editors of the People's Daily Online, and shared results of this thesis (translated into Chinese by the author) with them. Findings from this research investigation (such as information related to the influence of homepage positions on top news reports' pageview numbers) were highly appraised by the editors in Beijing and have played some positive role for the *People's Daily Online* to redesign its English and Chinese homepages. For the English edition, the "B" section used to be the place full of "the most popular news" from previous days, which were left there unchanged for up to seven days, and were "ignored" by the readers. The *People's Daily Online* redesigned its English homepage in 2009, and changed the "B"-section to a daily updated column. For the Chinese edition, the size of the Homepage Pic was reduced during the homepage re-designing process. And, the editors there also began to pay much more attention to making the titles attractive rather than competing for the better positions on the homepage. These encompass all the practical contributions made by this doctoral thesis to the booming Chinese Web information economy, especially the future development of the leading online newspaper.

This thesis is an interdisciplinary study in that it relates to both information science and journalism research. In such an era of globalization, more and more people rely on the Web to receive news. Methods and findings of this research are very helpful for news websites in China and potentially elsewhere around the world

(i.e., U.S. or Canadian international news agencies) to thoroughly understand the online readers' interests, as well as the relative geographic coverage of specific online news media, so as to keep current customers and attract new users, which are all crucial for the information industry's sustainable development in the coming years.

With the help of the thesis' findings on readers' comments, we could investigate readers' interests through a thorough content analysis of the comments posted to any type of reports about China or other nations. Studying such trends over a long period, we could see the changing of Chinese people's minds on some issues or attitudes towards other nations by analyzing the Web server logs of the leading news websites, such as the *People's Daily Online*. This type of study would potentially gather more objective data than surveys, especially when controversial or sensitive issues are involved. With a content analysis of readers' comments on the "hot" events, we may gain deeper insight into readers' interests. We could also address other issues in our future studies, such as whether the young and well-educated Chinese netizens are moderate nationalists or the so-called "populist Web mobs".

6.3 – Limitations of the Study

There are some limitations of the present thesis. First, the current study assumes that pageview number is a measure of reading interest and public mood. While this assumption proved to be valid for the finding that the results from the study of United States/Japan/Russia-related news reports paralleled the reality of

the public sentiments in China, it should be acknowledged that the pageview measure is still an indirect measure for other issues.

Second, the online survey is a direct measure to study readers' interests, however, there are always some human factors involved in the design of surveys, and participants of the online surveys only account for a small fraction of the whole number of visitors to a Website. For example, it is very easy for the editor to directly remove sensitive or controversial options before the survey is published online. This approach might significantly limit the quality of collected data, and might lead to uninformed conclusions.

Considering all of these limitations, no single method could draw a complete picture of online newsreaders' interests; various research methodologies, such as the survey, server logs and other publicly accessible Web information (e.g., comments posted by the readers) should be employed together, so as to measure readers' interests from different angles, and thus enable more comprehensive conclusions.

The server logs are too sensitive for the public to retrieve or share, but comment counts would play a somewhat positive role for us to study online readers' interests within the public domain, although there are also some concerns regarding the analysis of comments. First, the news editors in China routinely filter or withdraw "inappropriate" comments posted by the readers. The political contents on websites or BBS in China are censored. Any comment that dares to criticize the leaders or attack the CPC is removed immediately. Second, some readers might also intentionally leave fake or misleading information on the message bulletin board.

However, we could still gauge readers' interests or the public opinion via the comments, because these comments are the words from the real readers and a direct measure of their sentiments. And, leaving a message proves that they have at least browsed some content of the article, rather than just clicking on it or only viewing the title. The internal investigation like the one the author did, which was done with full information access to the un-censored data, will always outperform the work of outsiders' on "filtered comments". If the readers' comments are the only data available to conduct research on the public interest, we must note these limitations. We must also take precautions to deal with the limitations (i.e., the incompleteness of data) of inlinks and online reprints retrieved by commercial search engines.

6.4 – China's Future Web and the Need for More Research

The Web is a double-edged sword for the Chinese government and netizens. On the one hand, the growing strength and influence of the Web population prompted concern in Beijing about potential social unrest, and the government has stepped up its control over the Web in recent years. For example, China sends many "online police" to patrol the virtual space for "harmful activities and vulgar information" (e.g., online gambling, pornography, videos of protests, etc.). Some Web companies, such as Baidu and Google China, offered apologies to the public for the porn contents on their servers in January 2009 (People's Daily Online, 2009b). The administrative authority also asked the news sites to remove contents "unfit for the public to view" by themselves, especially those criticizing the current policies or officials during "politically sensitive periods". The BBS or the comment-

posting functions of the sensitive reports were closed, such as what happened during the tainted baby formula scandal in September 2008. Full access to Twitter, Facebook, and YouTube were blocked in China last March following unrest in Tibet (Hornby, 2009). More recently, after riots on July 5 in the capital of China's Xinjiang region, the local government strictly controlled Web access in most of that area to "maintain social stability" (Xinhua, 2009).

One the other hand, China's fast-growing online population has made the Web a forum for the Chinese people to express their opinions in a way rarely seen in the traditional media. For example, the Web rallied the Chinese people to protest online or on the street against the distorted reports on Tibet after the riots in Lhasa during March 2008. The Web also mobilized tens of thousands of disaster relief volunteers after the devastating Sichuan Earthquake in May 2008. The popular "citizen journalism movements" appeared in the Chinese Web space provided much valuable information to the public via the Blogs or even in the comments posted to the news reports.

However, due to the strict control and censorship of the Web by the Chinese authority, some political figures from the Western world, especially those in the United States, admitted that the Web alone cannot speed up the "democratization process" in countries like China in the near future (USCC, 2005). These political and social phenomena on the Web in China all merit study in the future.

In 2008, only 19% of China's population could access the Web, while 70% of the U.S. population was online at the same time. Another developed country

mentioned in this thesis' research questions and data analysis – Japan – enjoyed similarly high percentages (Barboza, 2008). Thus, China's Web market has strong potentiality for growth, especially in the online advertising business. The investment firm Morgan Stanley noted in Barboza (2008) that online advertising in China is growing by 60% to 70 % annually, and is expected to surge eight times to reach \$17 billion in 2016. The on-going global economic slump, which has affected all major business fields, may deter the rising trend of the Web market in China, but online newspapers must brace for the expected booming days after the crisis. Learning readers' interests and offering better-composed, timely updated news reports should always be their top priorities. Thus, possible future studies piloted by the presented thesis might yield more promising findings.

The Boston Consulting Group pointed out that China has a large number of mobile and Web users, but the penetration rate was still low, Xinhua (2008b) said, adding that Chinese Web users' habits differ from those of Westerners: Chinese people use online chats and text messages much more than Western people, who use e-mail as their major means of online communication. According to the CNNIC (2009a), China had already been the top of the world's leaders in cell phone use, and about 28.9% of this country's mobile telecommunication clients, or 73.05 million, browsed the Web via their cell phones, which had gradually replaced newspapers in commuters' hands.

With the development of the third generation (3G) mobile network, the wireless Internet could see explosive growth in the next few years (CNNIC, 2009b).

However, visits contributed by cell phones to the *People's Daily Online* accounted for less than 1 percent of total pageviews, although they had been increasing throughout 2008. So far, news reports offered to the "Cell Phoned Edition" are the same as those on the "PC-Editions", which need some major revisions. Due to the limitation of the screen size of the cell phone, news titles for these devices should be brief and attractive. New development of the "mobile newspaper" also requires more studies.

Nearly 70% of China's Web users were 30 or younger, and high school students were the fastest-growing group of new users, accounting for 39 million of the 43 million new netizens (CNNIC, 2009a). It is almost impossible for Web surfers at this age to read the parental *People's Daily*, which is too "old-fashioned" for their generation. If these Web users are ignored, this online newspaper will lose ground for future development. In addition to "selling" the *People's Daily Online* to students in the universities or colleges, this online newspaper must add more features to itself, such as video news and some space for young readers to express their opinions with relatively more freedom.

Of course, content of the news report is always the king. Publishing more exclusive stories and offering more comprehensive coverage of hot events than others will help the *People's Daily Online* establish a good reputation among the young "Web generation". If the editorial rules are too strict for online comments, Web users might migrate to other places to express their views. Unfortunately, there seems to be little room for the *People's Daily Online* to dramatically loosen control

over the readers' comments in the near future. How to attract more young readers also needs to be addressed in the future.

The findings from this thesis offered some practical knowledge to the biggest online newspaper in China, and more importantly, made some contributions to research methodology issues relevant to information science and media studies. The various ways of measuring readers' interests examined in this thesis could all be used and refined in future research. Server log statistics and publicly accessible data have both advantages and disadvantages to study readers' interests. The former is generally accurate but difficult to obtain, while the latter cannot directly reflect readers' interests and is less representative.

We must recognize that conclusions drawn from the server logs statistics as well as some publicly accessible data might reflect readers' interests only in part, as we only sampled a small portion of the Web users from some specific websites. Nevertheless, this doctoral thesis sheds some light on investigating public interest, and laid many useful foundations for future research.

References

- Ajiferuke, I., & Wolfram, D. (2004). Modelling the characteristics of Web page outlinks. *Scientometrics*, 59 (1), 43-62.
- Ajiferuke, I., Wolfram, D., & Famoye, F. (2006). Sample size and informetric model goodness-of-fit outcomes: a search engine log case study. *Journal of Information Science*, 32(3), 212-222.
- Alexa. (2009). About Alexa Internet. *Alexa Internet Inc.* Retrieved January 09, 2009, from: <http://www.alexa.com/company>.
- Almind, T. C., & Ingwersen, P. (1997). Informetrics analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4), 404 -426.
- AlShehri, F., & Gunter, B. (2002). The market for electronic newspapers in the Arab World. *Aslib Proceedings*, 54(1), 56-70.
- Ball, P. (2005). Life is short in online news. *Nature News*. Retrieved May 23, 2005, from: <http://www.nature.com/news/2005/050527/full/news050523-10.html>.
- Bar-Ilan, J. (1997). The "mad cow disease" Usenet newsgroups and bibliometric laws. *Scientometrics*, 39(1), 29-55.
- Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation bias is significant in determining user preference for search results - A user study. *Journal of the American Society for Information Science and Technology*, 60(1), 135-149.
- Barboza, D. (2008). China surpasses U.S. in number of Internet users. *New York Times*. Retrieved January 09, 2009, from <http://www.nytimes.com/2008/07/26/business/worldbusiness/26internet.html?scp=6&sq=china%20internet&st=cse>.
- BBC. (2008). Chinese government blocks overseas news websites again. *BBC*. Retrieved January 09, 2009, from: http://newsvote.bbc.co.uk/chinese/simp/hi/newsid_7780000/newsid_7785400/7785493.stm.
- Bjorneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1), 65-82.

- Bjorneborn, L., & Ingwersen, P. (2004). Towards a basic framework of webometrics. *Journal of the American Society for Information Science and Technology*, 1216-1227.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh WWW Conference*. Brisbane, Australia, April, 1998 (pp.107-117).
- Brown, C. (2004). Knowing where they're going: statistics for online government document access through the OPAC. *Online Information Review*, 28(6), 396-409.
- Chan, L. C. Y., Jin, B., Rousseau, R., Vaughan, L., & Yu, Y. (2003). Newspaper Coverage of SARS: A Comparison among Canada, Hong Kong, Mainland China and Western Europe. *Cybermetrics*, 6(1). Retrieved January 09, 2009, from <http://www.cindoc.csic.es/cybermetrics/articles/v6i1p1.html>.
- Chau, M. & Chen, H. (2003). Personalized and focused Web spiders. In: Zhong, N., Liu, J., Yao, Y. (Eds.), *Web intelligence*, Springer-Verlag. p. 197-217.
- Chau, M., Fang, X., & Sheng, O.R.L.(2005). Analysis of the query logs of a Web site search engine. *Journal of the American Society for Information Science and Technology*, 56 (13), 1363-1376.
- Chau, M., Lu, Y., Fang, X., & Yang, C., (2007). Web searching in Chinese: A study of a search engine in Hong Kong. *Journal of the American Society for Information Science and Technology*. 58 (7). 1044-1054.
- Chen, H., & Cooper, M.D. (2001). Using clustering techniques to detect usage pattern in a web-based information system. *Journal of the American Society for Information Science and Technology*, 52(11), 888-904.
- Chen, H., & Cooper, M.D.(2002). Stochastic modeling of usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology*, 53(7), 536-548.
- China Daily. (2008). China Internet users soar to 298 million. *China Daily*. Retrieved January 09, 2009, from http://www.chinadaily.com.cn/china/2009-01/14/content_7396500.htm
- China Daily. (2009). Hu calls on joint efforts to settle China-Japan disputes. *China Daily*. Retrieved May 09, 2009 from http://www.chinadaily.com.cn/china/2009-04/30/content_7735095.htm

- China Post. (2006). KMT shuts down Central Daily News after 78 years. *China Post*. Retrieved January 09, 2009, from <http://www.chinapost.com.tw/taiwan/detail.asp?ID=82855&GRP=B>
- CASS. (2002). *The National Survey of Web Use in China*. Beijing: Chinese Academy of Social Sciences (CASS). Retrieved January 09, 2009, from <http://www.people.com.cn/GB/it/8219/29993/index.html> (in Chinese).
- Chyi, H. I., & Lasorsa, D. (1999). Access, use and preferences for online newspapers. *Newspaper Research Journal*, 20(4), 2-13.
- CNNIC. (2009a). *The 23rd Statistical Report on the Internet Development in China*. Beijing: China Internet Network Information Center (CNNIC). Retrieved May 01, 2009, from <http://www.cnnic.cn/uploadfiles/pdf/2009/3/23/153540.pdf>
- CNNIC. (2009b). *The 24th Statistical Report on the Internet Development in China*. Beijing: China Internet Network Information Center (CNNIC). Retrieved August 08, 2009, from <http://www.cnnic.net.cn/html/Dir/2009/07/28/5644.htm>
- Cohen, L.B. (2003). A two-tiered model for analyzing library website usage statistics, Part 1: Web server logs. *Portal: Libraries and the Academy*, 3(2), 315-326.
- Cronin, B. (2001). Bibliometrics and Beyond: Some thoughts on web-based citation analysis. *Journal of Information Science*, 27(1), 1-7.
- Davenport, E., & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In: Cronin, B. & Atkins, H. B. (eds.). *The web of knowledge: a festschrift in honor of Eugene Garfield*. Metford, NJ: Information Today Inc. ASIS Monograph Series, 517-534.
- D'haenens, L., Jankowski, N., & Heuvelman, A. (2004). News in online and print newspapers: differences in reader consumption and recall. *New Media & Society*, 6(3), 363-382.
- Duy, J., & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *Journal of Academic Librarianship*, 32(5) 512-517.
- Encyclopedia-Britannica. (2009). Nanjing Massacre. *Encyclopedia-Britannica*. Retrieved: January 09, 2009, from <http://www.britannica.com/EBchecked/topic/402618/Nanjing-Massacre>.

- Etzioni, O. (1996). The World Wide Web: quagmire or gold mine. *Communications of the ACM*, 39(11), 65-68.
- Fagan, J. C. (2002). Use of an academic library Web site search engine. *Reference & User Services Quarterly*, 41(3), 244-252.
- Feuilherade, P. (2004). Online newspapers tempt readers. *BBC*. Retrieved January 09, 2009, from <http://news.bbc.co.uk/2/hi/technology/3767267.stm>.
- Fleisher, C. S., & Bensoussan, B. E. (2003). *Strategic and Competitive Analysis: Methods and Techniques for Analyzing Business Competition*. Upper Saddle River, New Jersey: Prentice Hall.
- Gao, Y., & Vaughan, L. (2005). Web hyperlink profiles of news sites: A comparison of newspapers of USA, Canada, and China. *Aslib Proceedings: New Information Perspectives*, 57(5), 398 – 411.
- Gao, Y. (2009). *Development Strategy for the People's Daily Online*. Report Submitted to the Editorial Board of the *People's Daily Online*.
- Google (2008). We knew the Web was big..... *The Official Google Blog*. Retrieved January 09, 2009, from <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>
- Greer, J., & Mensing, D. (2004). U.S. news Web sites better, but small papers still lag. *Newspaper Research Journal*, 25(2), 98-112.
- Gu, R., Zhu, M., Zhao, L., & Zhang, N. (2008). Interest mining in virtual learning environments. *Online Information Review*, 32(2), 133-146.
- He, Z., & Zhu, J.(2002). The ecology of online newspapers: the case of China. *Media Culture & Society*, 24(1), 121-137.
- Hope, B.G., & Li, Z.(2004). Online newspapers: the impact of culture, sex, and age on the perceived importance of specified quality factors. *Information Research*, 9(4). Retrieved August 09, 2008, from <http://informationr.net/ir/9-4/paper197.html>.
- Hornby, L. (2009). "Unafraid" China apparently fears YouTube. *Reuters*. Retrieved May 01, 2009, from <http://www.reuters.com/article/technologyNews/idUSTRE52N1VN20090324>.

- Huntington, P., Nicholas, D., Jamali, H., & Watkinson, A. (2006). Obtaining subject data from log files using deep log analysis: case study Ohio LINK. *Journal of Information Science*, 32 (4), 299-308.
- Hurst, M. (2001). Layout and language: Challenges for table understanding on the Web. In *Proceedings of the First International Workshop on Web Document Analysis*, Seattle, WA, 27-30.
- Jacobs, A. & Wang, J.(2008). Chinese urge anti-west boycott over Tibet stance. *The New York Times*. Retrieved January 09, 2009, from <http://www.nytimes.com/2008/04/20/world/asia/20iht20china.12156426.html?scp=3&sq=Jacobs%20%20China%20boycott&st=cse>
- Jacoby, J., & Laskowski, M. S. (2004). Measurement and analysis of electronic reserve usage: Toward a new path in online library service assessment. *Portal : Libraries and the Academy*, 4(2), 219-232.
- Jana, S., & Chatterjee, S. (2004). Quantifying Web site visits using Web statistics: an extended cybermetrics study. *Online Information Review*, 28(3), 191-199.
- Kahn, J. (2005). In rare legal protest, Chinese seek boycott of Japan goods. *The New York Times*. Retrieved: May 7, 2009 from http://www.nytimes.com/2005/04/09/international/asia/09beijing.html?_r=1&pagewanted=print&position.
- Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM SIGKDD Explorations*, 1-15.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., & Paatero, V. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks* [Special Issue on Neural Networks for Data Mining and Knowledge Discovery], 574-585.
- Lambert, F. (2008). *Rewriting the "Rules" of Online Networked Community Information Services: A Case Study of the mycommunityinfo.ca Model*. London, Ont.: Faculty of Graduate Studies, University of Western Ontario.
- Larson, R. R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. *ASIS 96*. Retrieved January 09, 2009, from <http://sherlock.berkeley.edu/asis96/asis96.html>.
- Lewandowski, D. (2008). A three-year study on the freshness of Web search engine databases. *Journal of Information Science*, 34(6), 817-831.

- Lewison, G. (2002). From biomedical research to health improvement. *Scientometrics*, 54, 179-192.
- Lewison, G. (2003). Beyond outputs: new measures of biomedical research impact. *Aslib Proceedings*, 55, 32-42.
- Lin, C.A., & Jeffres, L.W. (2001). Comparing distinctions and similarities across websites of newspapers, radio stations, and television stations. *Journalism and Mass Communication Quarterly*, 78(3), 555-573.
- MacLeod, C. (2009). China plans media empire to boost image. *USA Today*. Retrieved March 1, 2009, from http://www.usatoday.com/news/world/2009-02-18-chinamedia_N.htm
- Massey, B. L. (2004). Examination of 38 Web newspapers shows nonlinear storytelling rare. *Newspaper Research Journal*. 25 (3), 96-102.
- Massey, B. L., & Levy, M. R. (1999). Interactivity, online journalism, and English-language Web newspapers in Asia. *Journalism and Mass Communication Quarterly*. 76(1), 138-151.
- Min, D. (2008). Chinese online media plays bigger role in 2008 (in Chinese). *People's Daily Online*. Retrieved January 09, 2009, from <http://media.people.com.cn/GB/8587788.html>.
- Montgomery, A.L., & Faloutsos, C. (2001). Identifying Web browsing trends and patterns. *IEEE Computer*, 94-95.
- Nicholas, D., Huntington, P., Lievesley, N., & Withey, R. (1999a). Cracking the code: Web log analysis. *Online and CD-ROM Review*, 23(5), 263-269.
- Nicholas, D., Huntington, P., Williams, P., Lievesley, N., Dobrowolski, T., & Withey, R. (1999b). Developing and testing methods to determine the use of web sites: case study newspapers. *Aslib Proceedings*, 51(5), 144-154.
- Nicholas, D., Huntington, P., Lievesley, N., & Wastinson, A. (2000). Evaluating consumer Web site logs: case study The Times/Sunday Times Web site. *Journal of Information Science*, 26(6), 399-411.
- Nicholas, D., Huntington, P., & Williams, P. (2002). Evaluating metrics for comparing the use of Web sites: case study two consumer health web sites. *Journal of Information Science*, 28(1), 63-75.

- Nicholas, D., Huntington, P., Williams, P., & Dobrowolski, T. (2004). Re-appraising information seeking behavior in a digital environment. *Journal of Documentation*, 60(1), 24-43.
- Nicholas, D., Huntington, P., Jamali, H., & Tenopir, C. (2006a). What deep log analysis tells us about the impact of big deals: case study Ohio LINK. *Journal of Documentation*. 62(4). 482-508.
- Nicholas, D., Huntington, P., Jamali, H., & Watkinson, A. (2006b). The information seeking behaviour of the users of digital scholarly journals. *Information Processing & Management*, 42(5), 1345-1365
- Nicholas, D., Huntington, P., Jamali, H., & Rowlands, I. (2008). Viewing and reading behaviour in a virtual environment: The full-text download and what can be read into it. *Aslib Proceedings*, 60(3), 185-198.
- Ngowi, R. (2009). Christian Science Monitor has new look, new timing. *The Associated Press*. Retrieved: April 2, 2009, from: <http://www.msnbc.msn.com/id/29902792/>
- Payne, N., & Thelwall, M. (2007). A longitudinal study of academic webs: Growth and stabilisation. *Scientometrics*, 71(3), 523-539.
- OCLC (2002). *OCLC Web Characterization Project*. Retrieved January 09, 2009, from: <http://wcp.oclc.org/>.
- People's Daily Online (1999). China strongly condemns NATO bombing. *People's Daily Online*. Retrieved January 09, 2009, from: http://english.people.com.cn/english/199905/09/enc_990509001001_TopNews.html
- People's Daily Online (2001). Chinese fighter bumped by US military surveillance plane. *People's Daily Online*. Retrieved January 09, 2009, from: http://english.people.com.cn/english/200104/02/eng20010402_66544.html
- People's Daily Online. (2009a). About Us. *People's Daily Online*. Retrieved January 09, 2009, from <http://english.people.com.cn/90827/90828/index.html>
- People's Daily Online. (2009b). China vows to intensify online porn crackdown After shutting down thousands of sites. *People's Daily Online*. Retrieved from January 09, 2009, from <http://english.people.com.cn/90001/90776/90882/6587598.html>.

- Reid, E. (2003) .Using web link analysis to detect and analyze hidden web communities. In: Vriens, D.(ed). *Information and Communications Technology for Competitive Intelligence*, Hilliard, Ohio: Ideal Group Inc.
- Ross, N.C.M., & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science*, 51(10), 949-958.
- Rousseau, R. (1997). Sitations, an exploratory study. *Cybermetrics*, 1. Retrieved July 09, 2009, from <http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html>
- Rousseau, R. (2001). Evolution in time of the number of hits in keyword searches on the internet during one year, with special attention to the use of the word Euro. Retrieved July 09, 2009, from <http://users.pandora.be/ronald.rousseau/Euro.PDF>.
- Seidman, E. (2007). We are flattered, but... *Bing Community*. Retrieved January 09, 2009, from <http://blogs.msdn.com/livesearch/archive/2007/03/28/we-are-flattered-but.aspx>
- Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33(1), 6-12.
- Spink, A. (2002). Introduction to the Special Issue on Web research. *Journal of the American Society for Information Science and Technology*, 53(2), 65-66.
- Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3), 107-109.
- Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.
- Stone, M.L. (1999). Server logs help shape Web strategies. *Advertising Age's Business Marketing*, 84(1), 19-20.
- Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management*, 28(1), 1-2.
- Thelwall, M. (2001). Commercial Web site links. *Internet Research: Electronic Networking Applications and Policy*, 11(2), 114 – 124.

- Thelwall, M., Vaughan, L., & Bjorneborn, L. (2005). Webometrics, in B. Cronin (Ed.), *Annual Review of Information Science and Technology* (39) (p.81-135). Medford, NJ: Information Today, Inc.
- Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, 59(11), 1702-1710.
- Thelwall, M., Li, X., Barjak, F., & Robinson, S. (2008). Assessing the international Web connectivity of research groups. *Aslib Proceedings*, 60(1), 18-31.
- Thiel, S. (1998). The online newspapers: a postmodern medium. *The Journal of Electronic Publishing*, 4(1). Retrieved January 07, 2006, from www.press.umich.edu/jep/04-01/thiel.html.
- Tremayne M. (2004). The Web of context: Applying network theory to the use of hyperlinks in journalism on the Web. *Journalism and Mass Communication Quarterly*, 81(2), 273-253.
- Tsui, L. (2008). The hyperlinked society: questioning connections in the digital age. In J. Turow, & L. Tsui (Eds.), *The Hyperlinked Society: Questioning connections in the digital age* (pp.70-84). Ann Arbor: University of Michigan Press and University of Michigan Library.
- Tweddle, S., Avis, P., Davis, D., James, N., & Daniels, H. (1998). A method for investigating the usage of a cancer website. *New Technology in the Human Services*, 11(1), 12-16.
- USCC. (2005). *An analysis of the anti-American sentiments among some Chinese Internet users*. U.S.-China Economic and Security Review Commission (USCC). Retrieved January 09, 2009, from http://www.uscc.gov/researchpapers/2000_2003/reports/sentim.htm.
- Vaughan, L. (2001). *Statistical methods for the information professional: a practical, painless approach to understanding, using, and interpreting statistics*. Medford, N.J.: Information Today.
- Vaughan, L. (2004a). Web hyperlinks reflect business performance: A study of US and Chinese IT companies. *Canadian Journal of Information and Library Science*, 28 (1), 17-31.
- Vaughan, L. (2004b). Exploring website features for business information. *Scientometrics*, 61(3), 467-477.

- Vaughan, L. (2005). Web hyperlink analysis. In K. Kempf-Leonard (ed.), *Encyclopedia of social measurement* (pp.949-954). San Diego, CA: Academic Press.
- Vaughan, L. (2006). Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology*, 57(9), 1178-1193.
- Vaughan, L., & Hysen, K. (2002). Relationship between links to journal websites and Impact Factors. *ASLIB Proceedings*, 54(6), 356-361.
- Vaughan, L., Kipp, M., & Gao, Y. (2007a). Why are Web sites co-linked? The case of Canadian universities. *Scientometrics*, 72 (1), 81-92.
- Vaughan, L., Kipp, M., & Gao, Y. (2007b). Are co-linked business Web sites really related? A link classification study. *Online Information Review*, 31(4), 440-450.
- Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? *Journal of the American Society for Information Science and Technology*, 54(1), 29-38.
- Vaughan, L., & Zhang, Y. (2007). Equal representation by search engines? A comparison of websites across countries and domains. *Journal of Computer-Mediated Communication*, 12(3), 888-909.
- Wang, C. (2009). Attaching Great Significance to Translation, Increasing Global Communication. *Information Office of the State Council*. Retrieved November 26, 2009, from http://www.gov.cn/gzdt/2009-11/12/content_1462749.htm.
- Wang, P., Berry, M.W., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54(8), 743-758.
- Williams, P., & Nicholas, D. (1999). The migration of news to the Web. *Aslib Proceedings*, 51(4), 122-134.
- Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport, Conn.: Libraries Unlimited.
- Wolfram, D. (2008). Search characteristics in different types of Web-based IR environments: Are they the same? *Information Processing & Management*, 44(3), 1279-1292.

- Wolfram, D., Spink, A., Jansen, B., & Saracevic, T. (2001). Vox populi: The public searching of the Web. *Journal of the American Society for Information Science and Technology*, 52(12), 1073-1074.
- Wolfram, D., Wang, P., & Zhang, J. (2009). Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology*, 60(5), 896-910.
- Wolfram, D., & Xie, H. (2002). Traditional IR for Web users: A context for general audience digital libraries. *Information Processing & Management*, 38(5), 627-648.
- Wormell, I. (2001). Informetrics and webometrics for measuring impact, visibility, and connectivity in science, politics, and business. *Competitive Intelligence Review*, 12(1), 12-23.
- Wu, H.D., & Bechtel, A. (2002). Web site use and news topic and type. *Journalism and Mass Communication Quarterly*, 79(1), 73-86.
- Xinhua. (2008a). China, U.S. pledge constructive ties in future. *Xinhua News Agency*. Retrieved January 09, 2009, from http://news.xinhuanet.com/english/2008-06/29/content_8458631.htm.
- Xinhua. (2008b). TMT: Chinese internet users "huge potential e-commerce market". *Xinhua News Agency*. Retrieved January 09, 2009, from http://www.chinadaily.com.cn/bizchina/2008-07/10/content_6835092.htm.
- Xinhua. (2009). Xinjiang to Strictly control the Internet access and content. *Xinhua News Agency*. Retrieved October 09, 2009, from http://www.xj.xinhuanet.com/2009-09/24/content_17794484.htm
- Xinhua. (2009). China vows to build media's global communication capacity. *Xinhua News Agency*. Retrieved November 26, 2009, from <http://english.people.com.cn/90001/90776/90883/6825749.html>
- Xue, S. (2004). Web usage statistics and Web site evaluation: a case study of a government publications library Web site. *Online Information Review*, 28(3), 180-190.
- Zakaria, F. (2009). *The post American world*. New York: W.W. Norton & Company.
- Zhong, N., Liu J., & Yao, Y. (Eds.). (2003). *Web Intelligence*. New York: Springer.

Appendix 1: Classification Scheme for the Top 50 News Reports

1. Chinese Politics

This category of news reports includes those on the Chinese government's activities / policies / remarks on both domestic and world political events. Foreign nations' activities / policies / remarks on Chinese issues (including those on Hong Kong and overseas Chinese issues) are also included in this category. Chinese leaders' meetings with foreign counterparts, as well as China's stance on the situations of other nations, such as Iran or North Korea nuclear issues, are also included in this category.

Examples of topics: anti-corruption; army building; China's ties with other countries; all kinds of political policies; Party and government building, etc.

Chinese Report Example: 胡锦涛会见解放军和武警部队出席党的十七大

English Report Example: Hu Jintao meets Army delegates of Party Congress

2. World Politics

This category of news reports includes those on nations' (except China) movements / remarks / policies on their domestic and international political events (not including China-related issues); all kinds of regional wars; the activities of the United Nations.

Examples of topics: Iraq War; Georgian War; U.S.-Iran tensions; North Korea nuclear issues; Middle East peace process; civil wars/demonstrations/protests, etc.

Chinese Report Example: 拉登发布新录音带 要求伊拉克境内武装团结抗美

English Report Example: Bin Laden calls for more fighting against U.S. forces in Iraq

3. Business

This category of news reports includes those on business / economic policies and performance of China and other nations.

Examples of topics: governments' economic policies; stock markets fluctuations; state-owned enterprises and rural economic reforms; China's stock and housing market issues; the U.S. financial crisis; the rocket rising of oil prices; world market / business information; auto markets, etc.

Chinese Report Example: 证监会有关人士表示“AH 股份互换”系媒体错报

English Report Example: China has no plan to swap A, H shares

4. Taiwan

This category of news reports includes those on Taiwan's political and economic issues.

Examples of topics: Chinese mainland's stances/movements on Taiwan; the Taiwanese political figures' remarks; other nations'/international organizations' remarks/activities/policies on Taiwan; Taiwan's "presidential" election; Taiwan's economy; Taiwan's military maneuvers; Taiwan's culture, etc.

Chinese Report Example: 国台办：大陆方面绝不会吞下“台独”这个苦果

English Report Example: Beijing to take measures against "Taiwan independence"

5. Entertainment

This category of news reports includes those on entertainment issues.

Examples of topics: introduction to newly released movies and music; pop stars; movie stars; Academy Awards, etc.

Chinese Report Example: 组图：第26届中国电影金鸡奖各奖项揭晓

English Report Example: 26th Golden Rooster and Hundred Flowers Film Festival

6. Sports

This category of news reports includes all kinds of sports related stories (other than those on the Beijing Olympic Games).

Examples of topics: soccer games, NBA, World Cup, tennis, etc.

Chinese Report Example: 足球王国结束 64 年等待 2014 男足世界杯落户巴西

English Report Example: Brazil to host 2014 World Cup

7. Health

This category of news reports includes those on health care/medicine related issues.

Examples of topics: all kinds of diseases and treatments; beauty or body shaping tips; medicine or human body research findings, etc.

Note: News reports on the government's health care policies were grouped into this category. News reports relating to both science and health issues, such as new scientific/medical discoveries on health problems (methods to diagnose or cure cancer/serious diseases), were also included in this category.

Chinese Report Example: 美国超级病菌致死人数或超艾滋

English Report Example: Study: "superbug" may kill more Americans than AIDs

8. Sci-Tech-Edu

This category of news reports includes those on latest science/technology development, other than human body/health/medicine-related issues, as well as all education-related stories.

Examples of topics: space exploration programs, new technologies, new or extinct animal species, China's education reform and policies, introduction to other countries' education systems, career/job-seeking issues in China and abroad, etc.

Note: The government's education policies were grouped into this category.

Chinese Report Example: 我国首颗月球探测卫星嫦娥一号发射成功

English Report Example: China's first lunar probe Chang'e-1 blasts off

Chinese Report Example: 教育部：我国正举办着世界上最大规模的教育

English Report Example: 25 million students at university in China

9. Accidents/Natural Disasters

This category of news reports includes those on human-related accidents or natural disasters happening in China or other countries. Reports are placed in this category irrespective of the motivation of the accident (e.g., political/personal/uncontrollable technical reasons).

Chinese Report Example: 日本东京地铁突然停电 导致 9 万余人出行受影响

English Report Example: Subway blackouts disrupt rush hour in Tokyo

10. Odd

This category of news reports includes those on bizarre/weird events, and abnormal / extreme behaviors of people in their daily life.

Chinese Report Example: 离婚后妻子进屋捉奸被判“非法侵入住宅罪”

English Report Example: U.S. woman survives 19 hours in ocean waters

11. Crime

This category of news reports includes those on criminal activities.

Examples of topics: criminal cases; killing/hurting of innocent people; searching for and sentencing of criminal suspects, etc.

Note: stories on bribes taken by government officials and amendments to anti-crime laws were classified as Chinese/World Politics. Reports on “war crimes” and the “international tribunal” were categorized as World Politics.

Chinese Report Example: 组图:《中国经营报》女记者武真遇害案告破

English Report Example: Five women killed in clothing store shooting in Chicago

12. Environment

This category of news reports includes reports on environmental protection issues.

Example of topics: environmental protection (endangered animals, such as panda and the South China Tiger), climate changes, and pollution prevention issues, etc.

Note: all reports on the governments' environmental regulations were allocated to this group.

Chinese Report Example: 周正龙发现华南虎尸体?蔬果袋有望免费

English Report Example: U.S. mends image in climate change

13. Culture/ Life / Society

This category of news reports includes those on China and foreign countries' cultural and social life issues (non-political and non-economic events).

Example topics: all kinds of festival celebrations; unique characteristics of a country/region; new archeological discoveries from human cultural relics, etc.

Chinese Report Example: 图集: 2008 春运表情

English Report Example: Tens of thousands celebrate New Year in Jakarta, Indonesia

14. Earthquake

This category of news reports includes those on Sichuan Earthquake and all of the disaster relief and reconstruction works.

Chinese Report Example: 政治局常委会召开会议部署抗震工作 胡锦涛主持

English Report Example: President Hu says quake relief top priority

15. Tibet

This category of news report includes those on Tibet-related political /economic issues.

Example topics: Dalai and Tibet; Tibet's economic development

Chinese Report Example: 拉萨首次公开宣判打砸抢烧事件中部分犯罪案件

English Report Example: 30 sentenced in Lhasa riots

16. Olympics

This category of news reports includes those on the Beijing Olympic Games and other related events before and after the 2008 Games.

Example topics: Olympic Torch Relays; preparatory work for the Olympic Games; Olympic Games; etc.

Chinese Report Example: 孙海平详解刘翔退赛原因

English Report Example: Liu Xiang pulls out, coaches laud him fighter

17. Other

All kinds of reports that cannot be classified into the above mentioned categories are placed into this category.

Appendix 2: Classification Scheme for the Top 50 Queries

1.Chinese Politics

This category of queries includes those items related to the Chinese government's activities / policies / remarks on both domestic and world political events. Queries for foreign nations' activities / policies / remarks on Chinese issues (including those on Hong Kong and overseas Chinese issues) are also included in this category. Queries for Chinese leaders' meetings with foreign counterparts, as well as China's stance on situations of other nations, such as Iran or North Korea nuclear issues, are also included in this category.

Examples of topics: anti-corruption; army building; China's ties with other countries; all kinds of political policies; Party and government building, etc.

Note: If the Chinese Party/Government/Leaders' names are used as search queries, they will be grouped into this category. If Chinese political names and their foreign counterparts appear together, these queries will also be classified as "Chinese Politics".

Chinese Query Examples: 胡锦涛, 中国, 美国

Lookup news report: 胡锦涛会见美国国防部长盖茨

English Query Examples: Hu Jintao, China, U.S.

Lookup news report: President Hu Jintao meets Pentagon chief

2.World Politics

This category of queries includes those times related to other nations' (except China) movements / remarks / policies on their domestic and international political events (not including China-related issues); all kinds of regional wars; the activities of the United Nations, etc.

Examples of topics: Iraq War; Georgian War; U.S.-Iran tensions; North Korea nuclear issues; Middle East peace process; civil wars/demonstrations/protests, etc.

Note: Foreign governments'/leaders' names and international organization's names will be grouped into this category.

Chinese Query Examples: 拉登, 伊拉克

Lookup news report: 拉登发布新录音带 要求伊拉克境内武装团结抗美

English Query Examples: Laden, Iraq

Lookup news report: Bin Laden calls for more fighting against US forces in Iraq

3. Business

This category of queries includes those items related to business / economic policies and performance of China or other nations.

Examples of topics: governments' economic policies; stock markets fluctuations; state-owned enterprises and rural economic reforms; China's stock and housing market issues; the U.S. financial crisis; the rocket rising of oil prices; world market / business information; auto markets, etc.

Chinese Query Examples: A 股, 证监会

Lookup news report: 证监会有关人士表示“AH 股份互换”系媒体错报

English Query Examples: A shares, H share

Lookup news report: China has no plan to swap A, H shares.

4. Taiwan

This category of queries includes items related to Taiwan's political/economic events.

Examples of topics: Chinese mainland's stance/movements on Taiwan; the Taiwanese political figures' remarks; other nations'/international organizations' remarks/activities/policies on Taiwan; Taiwan's "presidential" election; Taiwan's economy; Taiwan's military maneuvers; Taiwan's culture, etc.

Chinese Query Example: 台独

Lookup news report: 国台办: 大陆方面绝不会吞下“台独”这个苦果

English Query Example: Taiwan independence

Lookup news report: Chinese mainland to take necessary measures against "Taiwan independence"

5. Entertainment

This category of queries includes those items related to news reports on entertainment figures and related issues.

Examples of topics: introduction to newly released movies and music; pop stars; movie stars; Academy Awards, etc.

Chinese Query Example: 金鸡奖

Lookup news report: 组图: 第 26 届中国电影金鸡奖各奖项揭晓

English Query Example: Golden Rooster

Lookup news report: 26th Golden Rooster and Hundred Flowers Film Festival

6. Sports

This category of queries includes those items related to sports news stories (other than those on the Beijing Olympic Games).

Examples of topics: soccer games, NBA, World Cup, tennis, etc.

Chinese Query Example: 世界杯

Lookup news report: 足球王国结束 64 年等待 2014 男足世界杯落户巴西

English Query Example: World Cup

Lookup news report: Brazil to host 2014 World Cup

7. Health

This category of queries includes those items related to stories regarding medical/health care-related issues.

Examples of topics: all kinds of diseases and treatments; beauty or body shaping tips; medicine or human body research findings, etc.

Note: Queries for the government's health care policies will be grouped into this category. Queries relating to both science and health issues, such as new scientific/medical discoveries on health problems (methods to diagnose or cure cancer/serious diseases), will also be included in this category.

Chinese Query Example: 超级病菌

Lookup news report: 美国超级病菌致死人数或超艾滋

English Query Example: superbug

Lookup news report: Study: "superbug" may kill more Americans than AIDS

8. Sci-Tech-Edu

This category of queries includes those items for the latest science/technology developments, other than human body health/medicine-related issues, as well as all education-related stories.

Examples of topics: space exploration program, new technologies, new or extinct animal species, China's education reform and policies, introduction to other countries' education systems, career/job-seeking issues in China and abroad, etc.

Note: Queries for education policies will be grouped into this category.

Chinese Query Examples: 嫦娥一号, 月亮女神

Lookup news report: 我国首颗月球探测卫星嫦娥一号发射成功

English Query Examples: Chang'e, China moon

Lookup news report: China's first lunar probe Chang'e-1 blasts off

9. Accidents/Natural Disasters

This category of queries includes those items for human-related accidents or natural disasters happening in China or other countries. Queries are placed in this category irrespective of the motivation of the accident (e.g., political/personal/uncontrollable technical reasons).

Chinese Query Examples: 东京地铁, 地铁停电

Lookup news report: 日本东京地铁突然停电 导致 9 万余人出行受影响

English Query Examples: Tokyo Subway

Lookup news report: Subway blackouts disrupt rush hour in Tokyo

10. Crime

This category of queries includes those items for criminal activities.

Examples of topics: criminal cases; killing/hurting of innocent people; searching for and the sentencing of criminal suspects, etc.

Note: queries for bribes taken by government officials and amendments to anti-crime laws will be classified as Chinese/World Politics. Queries for "war crimes" and the "international tribunal" will be categorized as World Politics.

Chinese Query Example: 武真

Lookup news report: 组图:《中国经营报》女记者武真遇害案告破

English Query Example: US shooting

Lookup news report: Five women killed in clothing store shooting in Chicago

11. Environment

This category of queries includes those items for environmental protection issues.

Example of topics: environmental protection (endangered animals, such as panda and South China Tiger), climate changes, and pollution prevention issues, etc.

Note: all governments' environmental regulations will be sent to this group.

Chinese Query Examples: 周正龙, 华南虎

Lookup news report: 周正龙发现华南虎尸体?蔬果袋有望免费

English Query Example: climate change

Lookup news report: U.S. mends image in climate change

12. Culture/ Life / Society

This category of queries includes those items for China and foreign countries' cultural and social life issues (non-political and non-economic events).

Example topics: all kinds of festival celebrations; unique characteristics of a country/region; new archeological discoveries from human cultural relics, etc.

Chinese Query Examples: 春节, 春运, 鼠年

Lookup news report: 图集: 2008 春运表情

English Query Examples: new year, carnival

Lookup news report: Tens of thousands celebrate New Year in Jakarta, Indonesia

13. Earthquake

This category of queries includes those items for the Sichuan Earthquake and all of the disaster relief and reconstruction works.

Chinese Query Examples: 汶川

Lookup news report: 汶川地震死亡人数已达 11921 人

English Query Examples: Sichuang Earthquake

Lookup news report: China's earthquake survivor visits U.S. Congress

14. Tibet

This category of queries includes those items for Tibet-related political/economic issues.

Chinese Query Examples: 拉萨骚乱, 达赖

Lookup report: 西藏自治区就拉萨极少数人打、砸、抢、烧破坏活动答记者问

English Query Examples: Lhasa, Tibet, Dalai Lama

Lookup report: Tibet regional gov't: Sabotage in Lhasa masterminded by Dalai.

15.Olympics

This category of queries includes those items for the Beijing Olympics Games.

Example topics: Olympics Games, Torch Relays.

Chinese Query Example: 开幕式

Lookup news report: 外媒评价北京奥运开幕式出奇一致

English Query Example: Liu Xiang

Lookup news report: Liu Xiang pulls out, coaches laud him fighter

16. Odd

This category of news reports includes those on bizarre/weird events.

Chinese Query Example: 变性美人

Lookup news report: 富豪女警兼职做妓女 12 大惊艳变性美人

English Query Example: cobra

Lookup news report: Trader's 115 cobras seized in Malaysia

17.Other All kinds of queries that cannot be classified into the above mentioned categories are placed into this category.

Appendix 3: 30 Negative News Reports on U.S.-Related Issues

Date	English Title	Original Chinese Title	Pageviews
Jan 01	The "U.S. century" is ending	日刊:美元帝国黄昏已然来临 美国世纪正在结束	89687
Jan 20	Six U.S. soldiers killed in Iraq	“基地”组织设下“地雷屋”炸死 六美军	95676
Feb 07	Economists: U.S. economy slides to recession	经济学家认为美经济滑向衰退 的可能性增大	68729
Feb 20	A Car "flies" 60m and crashes into Tree	美国一轿车变“飞机” 凌空滑翔 60 米撞到大树上	47229
Mar 01	Alleged U.S. rapist freed in Japan	涉嫌强奸日本少女的驻日美军 士兵被释放	76102
Mar 03	New Chinese nuclear submarine deployed to deter U.S. Navy	中国最新核潜部署台海 美军将 付更大代价	466080
Mar 06	Russian bomber flies over U.S. aircraft carrier again	俄轰炸机再次低空飞越美军航 母 美军战机拦截	527646
Mar 15	U.S. bomber crashes in Guam	美军 B-1B 远程轰炸机关岛出 事泄露军情	78094
Mar 16	Pentagon: the Iraqi War might be groundless	美“自扇耳光”:五角大楼称萨达 姆确实被冤枉了	708301
Mar 25	U.S. is not the only Superpower	美国: 从“一超”退向“首强”	128396
Apr 30	How does the U.S. spy on China?	解析布什如何对中国实行海陆 空全方位侦听	207032
May 08	US blackmails Myanmar after cyclone disaster	缅甸风暴灾难过后更可怕 美国 欲“趁火打劫”	328651
Jun 04	U.S. soldier rapes his 3-month- old daughter	美国前伊战士兵强奸 3 个月大 的女儿	200519
Jun 05	New secrets of Clinton's affair with Lewinski	克林顿和莱温斯基偷情时 希拉 里就在白宫	168727
Jun 06	US Air Force senior officials resign together	美国空军总参谋长和空军部长 同时辞职	236164
Jun 09	“Hawaii Kingdom” seeks independence from US	“夏威夷王国”声称要从美国 “独立”	601769

Jun 14	US Navy aircrafts collide, one dead	美海军“大黄蜂”与“虎”式战机凌空相撞一死两伤	248771
Jun 25	US female teacher has sex with eight students	美一中学惊现“禽兽”女教师 与8名学生发生性关系	952796
Jun 27	Iraqi Parliament member kills two US soldiers	伊拉克议员手持 AK-47 狂扫打死2名美军士兵打伤4人	133461
Jul 01	U.S. is the 22nd most flourishing country	外媒评全球最稳定繁荣国家 美国仅列第二十二位	148972
Jul 05	Russia will use nuclear weapons against U.S. if necessary?	俄必要时将使用核武自卫 目标指向美国?	56994
Jul 07	Bush "dancing" while waiting for McCain	美国总统布什等待麦凯恩不耐烦竟然跳舞解闷	339539
Jul 09	Why 17 U.S. girls became pregnant together	文汇论坛: 反思 17 名女生集体怀孕当未婚妈妈	700650
Aug 09	U.S. should change its nuclear weapon policy	美印核交易开绿灯 拥有最大核武库的美国应从良	508824
Aug 24	Seven weapons to defeat U.S. aircraft carrier	“七种武器”瘫痪航母(图)	309641
Sep 09	Woman who microwaved one-month-old daughter gets life imprisonment	美一妇人微波炉“烹”死满月亲生女被判终身监禁	203004
Sep 15	Lehman Brothers near bankruptcy	雷曼兄弟濒临破产	386396
Sep 16	U.S. trains collision kills 26	美火车相撞 26 人死亡 130 多人受伤	305488
Sep 18	China's capital management better than U.S.	卫报: 中国在资本管理方面比美国更胜一筹	687468
Sep 20	U.S. landlord peeps on female tenants for 20 years	美变态房东偷窥数百女房客 20 年 出租屋遍布摄像头	199507

Appendix 4: 30 Non-Negative News Reports on U.S.-related Issues

Date	English Title	Original Chinese Title	Pageview
Jan 01	US presidential election begins in Iowa	美国 08 大选: 参选人在爱奥华州首次较量	18736
Jan 20	Buffett reveals bailout plan for the market	财经人物: 巴菲特公布救市计划	11724
Feb 07	Super Tuesday: McCain and Hilary win	超级星期二: 麦凯恩大获全胜, 希拉里涉险过关	26505
Feb 20	US will shoot down a failed satellite soon	美国预计 21 日上午开展“导弹打卫星”活动	25889
Mar 02	Buffett plans to retire	“股神”巴菲特拟引退 将从四名人选中选接班人	23874
Mar 03	The running of US government: rule of law	美国政府运作有法可依	31920
Mar 06	Hilary get best gift for "Women's Day"	希拉里收到最佳三八节礼物	38698
Mar 15	US President's daughter will get married	布什要嫁女儿称是敏感外交	37221
Mar 16	Bush: Iraqi war will continue	布什将就伊战 5 周年发表讲话: 伊战将继续打下去	54806
Mar 25	"Brain dead" US man recovered	美脑死亡男子器官即将移植时苏醒 现已痊愈	43027
Apr 29	USS Kitty Hawk makes final port call in Hong Kong	美国航空母舰“小鹰号”抵港	105297
May 09	Who will be the female US vice president	盘点美国政坛巾帼: 谁将成为首位女副总统?(组图)	103344
Jun 04	Obama wins Democratic nomination	奥巴马获足够代表票 赢得民主党总统候选人提名	101507
Jun 05	China's earthquake survivor visits U.S. Congress	幸存汶川高中女生感动美国国会	579001
Jun 07	US Red-cross donates another 10 million to China	布什积极评价我赈灾工作 美红十字会再捐 1000 万	110978
Jun 10	Bush will try to catch Bin Laden before leaving White House	布什欲离任前抓住拉登	328006
Jun 12	Report: Obama will raise money in China	美媒披露: 奥巴马阵营将来中国筹款	45726
Jun 24	Bill Gates will donate all of his personal wealth	盖茨宣布捐出全部个人资产	109290

Jun 25	US will have the first female four star general	破例提拔 美或将诞生历史上第一位女性四星上将	100087
Jul 02	Why does Rice visit China now?	美国国务卿赖斯访华 目的何在?	166325
Jul 04	U.S. "Man" delivers baby girl	美国“男子”顺利生下健康女婴	251409
Jul 07	Bush will attend the Olympic opening ceremony	布什: 缺席奥运开幕式将是对中国人民的失敬	217953
Jul 13	US Senate approves 300b bailout plan for housing market	美参议院批准 3000 亿美元房地产救市计划	115654
Aug 06	US President visits South Korea	美国总统布什抵达韩国访问	72148
Aug 23	Obama chooses Biden as running partner	美媒:奥巴马选择拜登为其搭档竞选美国总统(图)	74193
Sep 09	Obama: I am thin but strong	奥巴马: 我虽瘦我强壮	73256
Sep 14	U.S. and Pakistan ties after Musharraf's resignation	穆沙拉夫下台了 美国怎么办?	100998
Sep 16	White people will not be the Majority in U.S. around 2050	美国种族结构多样化本世纪中叶变色	113462
Sep 18	If US and Iran resume ties, the world economy will change	美国伊朗重修旧好, 世界经济洗牌在即	203341
Sep 20	Gore endorses Obama	戈尔: 奥巴马是新林肯 麦凯恩走布什老路荒谬	72502

Appendix 5: 30 Negative News Reports on Japan-Related Issues

Date	English Title	Original Chinese Title	Pageview
Jan.15	Russia denies sending spy to Japanese PM office	福田办公室揪出俄间谍 俄坚称有人破坏两国关系	132380
Jan.27	Man attacks Yasukuni Shrine in Japan	男子靖国神社折断太阳旗 对日本人拳打	210245
Jan.31	Japan speeds up building missile defense system	日加速导弹拦截系统建设 再炒“中国威胁论”	173269
Mar.02	"Comfort Lady"(Japan's military sex slave in WWII) found in Hainan	海南老人王玉开公开“慰安妇”悲惨身世	162400
Apr. 28	Japanese female employee spends 350 thousand yen in one night	日本女白领一晚豪掷 35 万找男艺伎寻欢 (图)	139178
Apr. 30	China refutes Japan's view on "cross border pollution"	日本大米减产竟怨中国 专家驳斥日“越境污染说”	149547
May 01	Chinese student beaten by a Japanese right wing activist	圣火长野传递 中国留学生被日本右翼打伤	281924
May 04	Japanese Official demoted for browsing porn sites 780 thousand times	日本公务员浏览黄色网站 78 万次遭降职	157603
May 09	Man kills 7 in Tokyo	日本一“马路魔鬼”当街行凶致十七人死伤	231081
May 22	Japan's ruling Party wings try to increase share of power	日本自民党各大老为“后福田”布局 小泉活动频繁	277640
Jun. 01	Japanese star's sex scandal	日本惊现“艳照门” 女星松隆子性爱照流出(图)	241351
Jun. 11	Japan and Taiwan patrol boats collide near Diaoyu Island	日本海上保安厅船只与中国台湾渔船在钓鱼岛近海相撞	597765
Jun. 15	Japanese Navy Destroyer fails to launch missile during war games	日护卫舰在环太联合军演中发射导弹 出现哑弹	453294
Jun. 17	Taiwan, Japan police boats confront near Diaoyu Island	组图:台海巡巡防艇在钓鱼岛海域与日本巡逻艇对峙	614869
Jun. 18	Taiwan People protest over Japan's action near Diaoyu Island	台湾民众持续抗议日本蛮横行为 怒烧日本国旗	178397
Jun. 28	A 74-year-old Japanese porn star	日本 74 岁的色情明星 一场没有归路的堕落	652749

Jul. 10	14-year-old Japanese boy hijacks a bus	组图：日本一 14 岁男孩劫持大巴	123150
Jul. 19	Girl-eater executed in Japan	东京处决食女童恶魔 与秋叶原屠夫同为"御宅族"	813838
Aug. 09	The Japanese government near collapses	日本：福田内阁被逼上深秋绝路	165525
Aug. 30	70-year-old Japanese woman pretends to be younger for cheating.	日本七旬老姬冒充中年单身女 以色骗财锒铛入狱	305017
Sep. 13	Some 200 "Killing forecast" on Japan's Web	日网络现 200 余条“杀人预告”	168339
Sep. 24	New Japanese PM will not be a "troublemaker" to China?	香港明报：麻生无力向中国制造麻烦？	493635
Sep. 25	Five-year-old Japanese girl dies naked on street	日本一名 5 岁女童裸死街头	102189
Oct. 05	New Japanese Cabinet: a "Young Master Group"	共同社：日本在野党批评麻生内阁是“少爷集团”	205158
Oct. 07	Report: US may ship nuclear weapons to Japan in case of emergency	日美曾密约美军在“有事”时将核武器运进日本	104472
Oct. 11	Media: Japan's defense policy pays more attention to southwest	日媒：日本防卫重点开始向西南转移	123188
Oct. 20	Japan's PM furious at his defense ministry	福田发怒：防卫省“太不像话”	191196
Oct. 31	Japan's most expensive fighter got fire	日“最昂贵战机”起飞起火 曾被发现机身存在缺陷	159335
Nov. 20	China becomes Japan's new target	环球时报：关注中国 处处寻找出路 中国成为日本新目标	146892
Nov. 29	Japan's former vice defense minister arrested	日本前防卫事务次官被捕震撼日本政坛 福田大怒	150824

Appendix 6: 30 Non-Negative News Reports on Japan-Related Issues

Date	English Title	Original Chinese Title	Pageview
Jan. 15	Japan's Buried Treasures in SE Asia in WWII	无价宝藏 揭开二战中日本在东南亚藏宝秘密	55291
Jan. 27	What are Japan's advantages over China?	三张图片告诉我们日本比我们强在哪儿...	170634
Feb. 01	Japan's FM Calls for stronger ties with China	日本外相表示重视发展日中关系	21375
Mar. 02	Chinese, Japanese students dance together	组图：日本访华女高中生与中国学生携手翩翩起舞	94834
Apr. 29	Japanese people's view on work and employment	看看日本人的就业观	203927
Apr. 30	Japan closely watches China's military power	日本防卫省：我们“关注”中国军力	198476
May 02	Japan welcomes Chinese President's upcoming visit	非常欢迎胡锦涛主席访问日本	124386
May 07	Why Japan attaches great importance to President Hu's visit?	日本各方高度重视胡主席访问表明什么？	66054
May 09	Chinese President holds talks with Japanese PM	胡锦涛主席同日本首相福田举行会谈	63721
May 21	Japanese rescuers pay respect to Chinese victims	日本搜救队员对遇难者的默哀让人肃然起敬	445034
May 29	Japan studies radar data of the Wenchuan Earthquake	日本宇航研究机构分析汶川地震前后雷达数据	182933
Jun. 13	Japanese People show sympathy over China's earthquake	日本人民对中国的同情心高涨	397828
Jun. 16	Japan issues warning 10 seconds ahead of earthquake	日地震前 10 秒发出预警	310960
Jun. 17	Japanese think tank finishes strategy to deal with China	日本智库历时 2 年完成对华战略 预测中国 5 种未来	105941
Jun.18	Why Japan's earthquake caused less loss?	六大因素确保日本地震损失不重	157861
Jun.30	Japanese Navy destroyer ends visit to China	日舰结束访华 日媒表现令人惊讶	143870
Jul. 08	Japanese women apologize for WWII in Beijing	组图：日本妇人卢沟桥下跪谢罪	203473

Jul. 20	Japan's youngest Prince is two years old now	日本小王子满两岁 民众默 认皇位继承人(图)	145813
Aug.7	Why Japan's PM reshuffled cabinet?	解读福田内阁改组	109206
Aug.30	61 year old Japanese woman gives birth	日本 61 岁女性代女儿生子 创日本最高龄产妇纪录	697412
Sep.13	The prospects for the Sino-Japan relationship under Japan's new PM	日本政局变动与中日关系 (望海楼)	100866
Sep.24	Pieces of Paper and Japan's spirits	从几张纸看日本人的务实 精神	63241
Sep.25	Japanese Banks "Bottom Fishing" in Wall Street	日投行抄底华尔街 美国惊 呼: 日本人又回来了!	87417
Oct.05	Japanese PM: establish strategic mutual beneficial Japan-China relationship	福田表示将致力于构建日 中战略互惠关系	51287
Oct.07	Japan's Moon orbiter captures HD picture of the earth	日本"月亮女神"卫星拍摄 到高清晰地球影像	69659
Oct.10	Russia may help Japan become UN security council member	俄罗斯称拟帮助日本成为 联合国安理会成员	67470
Oct.19	Miss Japan 2008 debuts	组图: 2008 年最美"日本小 姐"亮相	59900
Oct.30	China's Moon orbiter and Japan's	中日探月工程 "嫦娥一号" 对比"月亮女神".	123781
Nov.19	Exploring Japanese people's spiritual foundation	学者: "自我软弱"造就日 本(图)	198061
Nov.29	Chinese Navy Destroyer arrives in Tokyo Bay	中国海军"深圳"号导弹驱 逐舰驶抵东京湾	56243

Appendix 7: 30 Negative News Reports on Russia-Related Issues

Date	English Title	Original Chinese Title	Pageview
Mar.06	Russian bomber expelled by U.S. fighters	俄罗斯战略轰炸机遭美军战机拦截	527646
Mar.16	Russia fails to send US satellite into orbit	俄发射美通信卫星未能进入预定轨道	61672
Mar.23	Some Russian billionaires flee the nation to escape assassinations	俄多名流亡英国寡头担心遭暗杀秘密潜逃	47924
Mar.25	Russia postpones deadline of finishing East Oil Pipeline	俄工业与能源部推迟东线石油管道完工期限	91398
Mar.30	Russian cult members hide in cave for "Judgment Day"	俄罗斯邪教徒为躲避世界末日藏身洞穴	37987
Apr.08	Putin hints invading Ukraine over the latter's NATO membership	媒体称普京暗示乌克兰若加入北约将并吞乌领土	52857
Apr.13	Two Russian soldiers sell stolen tank for scrap steels	俄罗斯两名士兵盗窃坦克当废铁卖获刑	61680
Apr.17	Russian media: Putin divorced and to marry young girl	俄媒体称普京已经秘密离婚 将娶性感女运动员	76644
Apr.27	Georgia vows to retaliate Russia's invading	格鲁吉亚警告称将报复俄罗斯侵略行为	52551
May 01	Russia: corruption is a serious problem	俄官员称腐败吞噬三分之一财政预算	67890
May07	Georgia and Russia close to "War Status"	格鲁吉亚称格方与俄罗斯接近战争状态	98466
May18	Georgia Presses Russia over WTO entry talks	格鲁吉亚借入世谈判向俄罗斯施压	85413
May 27	Russian Plane crashed, 9 died	俄罗斯一架安-12 运输机坠毁 9 人遇难	166267
Jun. 01	Some hostage survivors wants to prosecute Putin	别斯兰人质事件幸存者要求审讯普京	94372
Jun.11	U.S. "extremely upset" with Russia's Domestic Politics: Rice	赖斯撰文称美国对俄国内政治发展极度失望	98673
Jun.12	Russia's ruling party will expel 30K members	普京下达清党指示 俄执政党将开除 3 万党员	107552

Jun.18	Georgia arrests 4 Russian Peacekeepers	格鲁吉亚扣留 4 名俄罗斯维和军人	120828
Jun.25	US: Russia blackmails Mongolian economy	美国媒体称蒙古遭俄罗斯经济敲诈	102135
Aug.04	Russian President and PM at odds	梅德韦杰夫首次与普京唱反调	192706
Aug.09	Five Russian planes shot down	格军在南奥激战数十人阵亡 称击落 5 架俄战机	601331
Aug.10	Russia: 2000 people died in South Ossetia	俄罗斯称南奥塞梯激战已致 2000 人死亡(图)	252621
Aug.11	Three Russian soldiers killed in South Ossetia	南奥塞梯称格方再次开火 3 名俄罗斯士兵丧生	179127
Aug.12	Bush: Russia "invading" Georgia Unacceptable	布什称美国无法接受俄“入侵”格鲁吉亚	106830
Aug.13	Georgia sues against Russia at the International Tribunal	格鲁吉亚向国际法院起诉俄罗斯	195035
Aug.14	Oil prices rise dramatically due to Russia-Georgia conflict	俄罗斯和格鲁吉亚战事导致世界油价疯狂上涨	157495
Aug.15	Bush: Russia violates truce agreement	布什称俄罗斯违反停火协议	205071
Aug.24	Russian armor vehicle attacked, 2 dead	俄军装甲车在车臣遇袭 2 人死亡 2 人受伤	172947
Aug.28	Georgia will cease diplomatic ties with Russia	格鲁吉亚准备与俄罗斯单方面断绝外交关系	155874
Aug.30	Russian, U.S. navy confront in the Black Sea	美俄战舰黑海对峙可能引发走火事件	168868
Aug.31	U.S. Russian passenger planes narrowly escape collision over the Caribbean Sea	美俄客机在加勒比海上空险些相撞	100295

Appendix 8: 30 Non-Negative News Reports on Russia-Related Issues

Date	English Title	Original Chinese Title	Pageview
Mar. 06	Russia needs strongman leader: Survey	调查称 42%俄民众认为国家需要斯大林式领袖	474210
Mar. 16	Assassination plot against Putin foiled	俄媒体称特工挫败一起针对普京暗杀阴谋	701921
Mar. 21	Russia, Japan will develop E. Serbia oilfield together	日俄协定共同开发东西伯利亚油田	48306
Mar. 22	Russia to negotiate with other nations manufacturing AK-47	俄拟与 AK-47 生产国就专利使用进行谈判	65463
Mar. 27	"Stalin Dumpling " popular in Russia	斯大林牌水饺畅销俄罗斯	148972
Apr. 07	US, Russia may build the Missile Defense System together	普京与布什考虑美俄欧共建导弹防御系统	72293
Apr. 13	Ukraine's NATO membership will lead to military reactions	俄官员称将以军事行动应对乌克兰加入北约	66119
Apr. 17	Putin elected Russia's ruling party leader	普京正式当选统一俄罗斯党主席任期四年(组图)	84032
Apr. 24	Interview with widow of former Russian President	专访叶利钦遗孀: 不能一辈子住国家别墅	73455
May 02	Putin only brings one pen from his Presidential Office	普京称卸任时将仅从克里姆林宫带走 1 支钢笔	229545
May 07	Putin to become Russian PM	俄国家杜马明日将任命普京为政府总理	113361
May 21	Russian billionaire spends 33.6m dollar on painting	俄罗斯首富以 3360 万美元购得“裸肥女”名画	115298
May 27	Russian President signs act to increase birth rates	俄罗斯总统签署政令鼓励生育	198533
May 28	Russia warns Ukraine and Georgia not join NATO	俄罗斯警告乌克兰和格鲁吉亚不要加入北约	110865
Jun. 10	Russia opposes military actions against Iran	俄罗斯反对武力解决伊朗核问题	121896
Jun. 12	Russia overtakes Saudi as the biggest oil producer	俄罗斯超过沙特成为最大产油国	189520
Jun. 18	Russia, U.S. call on cracking down against nuclear terrorism	俄美呼吁全世界联合打击核恐怖主义	159456

Jun. 25	Russia condemns some nations' rewriting of WWII history	俄总统谴责前苏联加盟国欲重写二战史做法	87964
Aug. 04	Russia sells 40 billion U.S. Bonds	俄罗斯抢先抛出美债券400 亿美元	1092899
Aug. 09	Russia sends armored troops to Georgia.	俄罗斯向冲突地区派装甲部队 格鲁吉亚军事行动升级	506008
Aug. 10	Russia bombs Georgian military facilities	格鲁吉亚：俄军空袭格军用设施	703521
Aug. 11	Russian Navy fires at Georgian boat	俄海军在阿布哈兹水域击沉一艘格鲁吉亚导弹快艇	426888
Aug. 12	Russia controls most of South Ossetia	俄海陆空三军齐攻格鲁吉亚 已控制南奥塞梯大部	209820
Aug. 13	Russia ends military actions in Georgia	梅德韦杰夫宣布俄结束对格鲁吉亚军事行动	471800
Aug. 14	Many tanks abandoned by Georgian Army fighting Russians	俄格军事冲突大批格军陷入恐慌弃坦克逃跑	323205
Aug. 15	Russian Special Forces blitzed Georgia	俄空降兵闪电切割格鲁吉亚	361788
Aug. 23	Russia to provide defensive weapons to Syria	俄罗斯将向叙利亚提供防御性武器	106267
Aug. 28	Russia recognizes the Independence of South Ossetia	俄总统签署法令承认南奥塞梯和阿布哈兹独立	396506
Aug. 29	Russian Black Sea Fleet anchors in the Port near South Ossetia	俄罗斯黑海舰队部分军舰驶入阿布哈兹海港	286127
Aug. 31	U.S. asks Russia to return captured Hammvees	媒体称美可能要求俄罗斯归还被缴获悍马	151538

Appendix 9: Type of Website

BBS Sites (BBS)

Example: A link from <http://forum.defence.org.cn/> to english.people.com.cn. The former is a BBS (online forum) website.

Corporate Websites run by companies (Com)

Example: A link from <http://www.explorenepal.com/about.php> to english.people.com.cn. The former is a corporate Website run by a company providing Web services.

Non-profit organization Websites (Org)

Example: A link from <http://www.clii.com.cn/> to www.people.com.cn. The former is a non-profit organization's website.

Government Websites (Gov)

Example: A link from <http://www.realestate.cei.gov.cn/> to www.people.com.cn. The former is a government website.

Weblogs (Weblogs)

Example: A link from <http://www.eurotrib.com/> to english.people.com.cn. The former is a Weblog.

Personal Websites (Websites maintained by individuals and the site content is of personal interest)

Example: A link from <http://www.hardcoreware.net> to english.people.com.cn. The former is a personal site.

News Websites (Website's main content is news reports)

Example: A link from <http://news.google.com/> to english.people.com.cn. The former is a news site.

Website run by News Media

Example: A link from <http://www.xinhuanet.com/mil/> to www.people.com.cn, the former is a News Media Website.

Educational Websites (Edu)

Example: A link from www2.cddc.vt.edu/ to www.people.com.cn. The former is an educational site.

Commercial Portals such as Yahoo! (Online Portal)

Example: A link from <http://www.bolaa.com/> to www.people.com.cn. The former is a commercial Web Portal.